Juha Kilpiäinen

TEKOÄLYN HAAVOITTUVUUDET KYBERTOIMINTAYMPÄRISTÖN NÄKÖKULMASTA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

TIIVISTELMÄ

Kilpiäinen, Juha
Tekoälyn haavoittuvuudet kybertoimintaympäristön näkökulmasta
Jyväskylä: Jyväskylän yliopisto, 2020, 80 s.
Kyberturvallisuus, pro gradu -tutkielma
Ohjaaja: Lehto, Martti

Nyky-yhteiskuntaamme on tullut olennainen lisä, tekoäly. Se koskettaa useita yhteisömme eri osia ja on tullut tukemaan erityisesti modernia palveluyhteiskuntaamme. Päivittäin modernin palveluyhteiskuntamme useat eri palvelut hyödyntävät tekoälyjärjestelmiä. Näitä ovat muun muassa tietoliikenne- pankki-, terveys- ja kaupanala sekä sähköinen viestintä. Sähköisessä viestinnässä erityisesti puhelin, televisio, radio, sosiaalinen media ja internet hyödyntävät tekoälyä. Esillä olevien positiivisten hyötyjen lisäksi tekoälykontekstissa on alkanut näkyä myös negatiivisia vaikutuksia. Tekoäly kuten mikään muukaan ICT-järjestelmä ei ole vastustuskykyinen toimijoille, jotka haluavat hyödyntää sen haavoittuvuuksia omien tavoitteiden saavuttamiseksi. Tekoälyn kyberturvallisuusympäristössä toimijan on ymmärrettävä, että tekoälyllä on haavoittuvuuksia, joita voidaan käyttää sitä vastaan. Tämä tekoälyn kielteinen hyödyntäminen koskettaa niin julkispalveluita, suuryrityksiä kuin yksilöitä.

Tässä pro gradussa tutkitaan tekoälyn haavoittuvuuksia. Tutkimus on rajattu koskemaan tekoälyn haavoittuvuuksia, jotka esiintyvät sen kybertoimintaympäristössä. Tutkimuksen kohdetta on syvennetty tarkastelemalla kybertoimintaympäristössä esiintyviä hyökkäyksiä tekoälyä vastaan.

Tutkimus on suuntautunut laadullisesti. Koska tutkimuskysymykseen liittyvästä ilmiöstä on vain vähän teoriaa ja tutkimuskirjallisuutta, aineistonkeruumenetelmänä käytettiin erilaisiin dokumentteihin, raportteihin ja julkaistuun kirjallisuuteen perustuvaa tietoa.

Johtopäätöksenä tekoälystä löytyy haavoittuvuuksia oppimiseen, opettamiseen, koulutusmateriaaliin, algoritmeihin, tunnistustekniikoihin ja ymmärrykseen liittyen. Haavoittuvuuksia on myös koodi-, ohjelmisto- ja laitetasolla sekä tietoverkoissa. Lisäksi yhä monimutkaisemmaksi muuttuvat tekoälyjärjestelmäkokonaisuudet korostavat näistä koostuvien ratkaisujen haavoittuvuutta. Tutkimuksessa esille tuodut tekoälyn haavoittuvuudet ovat vain haasteita, jotka ymmärrämme tekoälystä tänään. Teknologioiden kehittyessä ja uusien käyttötarkoitusten ilmaantuessa ei ole vaikea kuvitella tulevia uusia haavoittuvuuksia. Ne voivat jalostua esiin näistä aiheista tai ilmaantua muualta.

Asiasanat: tekoäly, haavoittuvuus, kybertoimintaympäristö, kyberturvallisuus

ABSTRACT

Kilpiäinen, Juha
Vulnerabilities in artificial intelligence from the perspective of the cyber-environment
Jyväskylä: University of Jyväskylä, 2020, 80 pp.
Cyber Security, Master's Thesis
Supervisor: Lehto, Martti

The everyday life of our modern society has changed considerably because of technological developments. In particular, artificial intelligence has become a significant factor in almost every sector of our society, including the sectors providing services for needs of society. In our daily life several different sectors of society and services utilize artificial intelligence systems. In the social sectors and services, such as telecommunications, banking, health and commerce, as well as telephone, television, radio, social media and the Internet, artificial intelligence plays a key role. In addition to the great benefits of artificial intelligence to our lives, negative effects have also begun to show up. Artificial intelligence, similar to ICT system, is not resilient to malicious actors who want to exploit its vulnerabilities to achieve their own goals. When operating in the cyber security environment of artificial intelligence, the actor must understand that artificial intelligence has vulnerabilities that can be exploited against it. This harmful use of artificial intelligence affects public services, large companies and individuals.

This Master's Thesis investigates the vulnerabilities of artificial intelligence. The research has been limited to the vulnerabilities of artificial intelligence that occur in the cyber-environment. The subject of the study has been deepened by to look into attacks against artificial intelligence in the cyber security environment.

The research is qualitatively oriented. As there is only a theory and research literature on the phenomenon available. Information was gathered from different sources such as various documents, reports and published literature.

In conclusion, there are vulnerabilities in artificial intelligence relating to learning, teaching, training materials, algorithms, recognition techniques and content understanding. There are also vulnerabilities in respect of codes, software, and hardware. as well as information networks. Moreover, the increasingly complex artificial intelligence system assemblies highlight the vulnerability of the solutions composed. The vulnerabilities of artificial intelligence discussed in the research are just challenges that we have found out about today´s artificial intelligence. As technologies evolve and new applications emerge, it is not difficult to imagine future new vulnerabilities. They can be refined from these subjects or they can appear out of nowhere.

Keywords: artificial intelligence, vulnerability, cyber environment, cybersecurity

SISÄLLYS

1 JOHDANTO

Tässä pro gradussa tutkitaan tekoälyn haavoittuvuuksia. Tutkimuksessa keskitytään siihen, millaisia tekoälyhaavoittuvuuksia on olemassa erityisesti kybertoimintaympäristössä. Tutkimusta täydennetään tutkimalla hyökkäyksiä tekoälyä vastaan. Koska hyökkäyksien kautta on mahdollista tutkia laajemmin tekoälyn eri haavoittuvuuksia, ne avartavat tutkimusongelmaan perehtymistä.

Vähäkainun, Lehdon ja Kariluodon (2020) mukaan termi tekoäly on ollut pinnalla jo vuosikymmeniä. Se esitettiin aluksi ihmisen aivojen kognitiivisten toimintojen jäljittelemiseksi. Tekoälyn keskeinen asia on, että se voi käsitellä merkittävän määrän tietoa. Tätä kautta sillä on uusia sovelluksia nykyisessä ympäristössämme (Vähäkainu, Lehto & Kariluoto, 2020). Nykyinen digitaalinen ympäristömme ja jokapäiväinen toimintamme on omaksunut tekoälyn yhdeksi palveluyhteiskuntamme osaksi. Siukosen ja Neittaanmäen (2019) mukaan moderni palveluyhteiskuntamme hyödyntää tekoälyä osana tätä digitalisaatiota. Tekoälyä hyödyntävät useat eri yhteiskunta-alat ja palvelut. Näitä ovat muun muassa tietoliikenne, pankki-, terveys- sekä kaupanala. Erityisesti tekoälyn vaikutus näkyy älylaiteteknologiassa, sosiaalisessa mediassa, internetissä ja televisiossa. Tekoälyä käytetään konkreettisesti esimerkiksi hakukoneissa, kasvojentunnistuksessa, ääniohjauksessa, kohdennetussa mainonnassa, roskapostin suodatuksessa sekä rekisteritunnistusjärjestelmissä ja niin edelleen (Siukonen & Neittaanmäki, 2019). Samaa mieltä Siukosen ja Neittaanmäen kanssa on myös Lehto. Lehdon (2019) mukaan tekoäly kehittyy huimin harppauksin. Tekoäly tulee mahdollistamaan eritoten ihmistä vaativien tehtävien automatisoinnin. Tämä kehitys tulee avaamaan uusia mahdollisuuksia uudistaa nykyisiä toimialoja. Tekoälyn kyvykkyys tulee painottumaan esimerkiksi analyyseissa ja havaintojen läpikäynnissä. Tämä siksi, koska tekoäly kykenee käsittelemään nopeasti valtavat määrät dataa ja luomaan tästä uutta tietoa. Jo nykyisin valtavien datamäärien käsittelemiseen tarvitaan älykkyyttä ja koneita. Tekoälyn avulla koneet, laitteet, järjestelmät, ohjelmat ja palvelut toimivat tilanteen mukaisesti. Samalla tekoäly tekee päätelmiä käytössään olevan tiedon perusteella ja avustaa ihmistä (Lehto, 2019). Vähäkainun ym. (2020) mukaan data on ensiarvoisen tärkeää toimivalle tekoälylle. Tekoälyn käyttämien tietojen pitää olla saatavilla ja oikeita (Vähäkainu, Lehto &

Kariluoto, 2020). Amodei ym. (2016) ja Fralick (2019) lisäävät, että datan määrän kasvaessa riippuvuus tekoälystä kasvaa.

Lähes kaikessa julkaistussa tekoälykirjallisuudessa ja raporteissa tekoäly nähdään mielenkiintoisena ja ajankohtaisena tutkimusalueena. Tämän allekirjoittavat muun muassa Mitchell (2018) ja Kilpatrick (2019b). Heidän raporttiensa mukaan tekoäly on yksi tärkeimmistä ja aktiivisimmista tutkimusalueista tietotekniikan alalla. Se on siirtynyt keskuuteemme tieteiskirjallisuuden alueelta. Raporttien mukaan olemme tekoälyn kanssa vuorovaikutuksessa päivittäin. Tähän vuorovaikutukseen sisältyy valitettavasti myös toinen puoli. Kuten kaikki tehokkaat tekniikat, tekoälykin on kaksiteräinen miekka. Tekoälyllä on potentiaalia parantaa elämäämme, mutta se tarjoaa myös kyberturvallisuuden haasteita (Mitchell, 2018 ja Kilpatrick, 2019b).

Băjenescun (2018) mukaan tekoälyn vallankumous on nopein kaikista tunnetuista vallankumouksista. Sen esitetään pystyvän parantamaan ja hienosäätämään jo olemassa olevia prosesseja. Sen sanotaan olevan paljon enemmän kuin tekniikka. Usein todetaankin, että tekoäly edustaa uutta tapaa olla vuorovaikutuksessa ympäristön ja liiketoiminnan kanssa. Useat suhtautuvat myönteisesti tekoälyinnovaatioon. Toiset vaativat hypetykseen taukoa, arvioidakseen kuluttajille mahdollisesti aiheutuvia riskejä tai haitallista toimintaa (Băjenescu, 2018). Stephensonin (2018), Mitchellin (2018) ja Kilpatrickin (2019b) mukaan on väistämätöntä, että riskejä esiintyy. Ennemmin tai myöhemmin tehokasta tekniikkaa käytetään haitallisiin tarkoituksiin. Useat tutkijat ja verkkoturvallisuuden ammattilaiset ovat antaneet hälytyksen tekoälyn käytöstä esimerkiksi perinteisten verkkoturvajärjestelmien heikentämiseksi (Stephenson, 2018; Mitchell, 2018 ja Kilpatrick, 2019b).

Mitchellin ja Kilpatrickin kanssa samaa mieltä ovat Amodei ym. (2016) ja Fralick (2019). Heidän mukaansa tekoälyn nopea kehitys on nostanut esiin tekoälytekniikoiden negatiiviset yhteiskunnalliset vaikutukset. Nämä voivat olla tahattomia tai tahallisia. Tekoälyn käytön kehittyessä ja lisääntyessä pahantahtoiset toimijat lisäävät tietämystään ja kykyjään hyödyntää tekoälyä (Amodei ym., 2016 ja Fralick, 2019). Jos tekoälyn taustalla olevan tekniikan tehostuminen jatkuu, lisätään haitalliseen käyttöön vain uusia keinoja täydentävät Shevlane ja Dafoe (2020).

Amodein ym. (2016) ja Fralickin (2019) mukaan kyberturvallisuuden alalla pitäisi hyväksyä se käsitys, että pöydän toisella puolella olevat ovat yhtä älykkäitä kuin mekin. He ymmärtävät, jos otamme riskejä. Tämän takia on noussut esille laaja ja monipuolinen keskustelu tekoälyyn liittyvistä kysymyksistä. Keskustelu pitää sisällään muun muassa onnettomuudet, etiikan, riskiherkkyyden ja turvallisuuden. Koska tekoälyjärjestelmiä käytetään yhä suuremmassa mittakaavassa, erilaisissa tilanteissa on syytä pohtia negatiivisten asioiden skaalautuvuutta. Skaalautuvuuden lisäksi on pohdittava sitä, mitkä haasteet on kohdattava onnettomuuksien riskien vähentämisessä (Amodei ym., 2016 ja Fralick, 2019). Amodei ym. (2016), Fralickin (2019) sekä Vähäkainun ym. (2020) mukaan, jos löydämme uuden uhan tai keksimme miten suojautua siltä, toiset suunnittelevat tapoja kiertää tai häiritä näitä. Amodein ym. (2016) ja Fralickin (2019) mukaan voimme olla varmoja siitä, että vastustajat hyödyntävät haavoittuvuuksia. Tekoäly ei ole poikkeus (Amodei ym., 2016 ja Fralick, 2019).

Suomessakin on tehty tekoälyyn liittyvää raportointia. Tekoälyohjelman loppuraportin (2019) mukaan tekoälystä on puhuttu viimeisten vuosien aikana laajasti. Laskentakapasiteetin tehostuminen, tekniikan halventuminen, datamäärän kasvu ja tekoälyalgoritmien kehittyminen, ovat johtaneet tekoälyn lisääntyneeseen hyödyntämiseen. Tekoälyyn on viitattu, kun on puhuttu tämän vuosisadan tärkeimmistä teknologioista. Tähän liittyen tekoälyohjelman loppuraportti korostaa tekoälyn olevan enemmän kuin vain yksi teknologia. Tekoäly on joukko erilaisia menetelmiä, teknologioita, sovelluksia ja jopa tutkimussuuntia. Nämä tekoälyn menetelmät, teknologiat ja sovellukset ovat osa digitalisaation laajempaa ilmiötä ja kehitystä (Tekoälyohjelman loppuraportti, 2019). Tätä tekoälyohjelman loppuraporttia täydentää Ailiston, Heikkilän, Helaakosken, Neuvosen ja Seppälän (2018) raportti: Tekoälyn kokonaiskuva ja osaamiskartoitus. Heidän mukaansa tekoälykenttää voi kuvata ja jäsentää esimerkiksi seuraavan jaottelun mukaisesti: data-analyysi, havainnointi, tilannetietoisuus, luonnollinen kieli, kognitio, vuorovaikutus ihmisen kanssa, digitaidot työelämässä, ongelmanratkaisu, laskennallinen luovuus, koneoppiminen, järjestelmätaso, systeemivaikutukset, tekoälyn laskentaympäristöt, alustat, palvelut, robotiikka ja koneautomaatio (Ailisto, Heikkilä, Helaakoski, Neuvonen & Seppälä, 2018). Tekoäly on siis kokonaisuus ja luonnollinen jatkumo, ei vain yksittäinen digitalisaation osa.

Tekoäly tuo paljon positiivista kehitystä, mutta sillä on myös varjopuolensa. Castelluccion (2018a) ja Kilpatrickin (2019) mukaan digitaalisessa tilassa ideoita käytetään sekä positiivisiin, että negatiivisiin tarkoituksiin. Tekoäly ei ole immuuni käyttäjille, jotka haluavat vahingoittaa sillä muita tai hyödyntää sen haavoittuvuuksia (Castelluccio, 2018a ja Kilpatrick, 2019). Kilpatrickin (2019) mukaan kyberturvallisuuden on käännettävä painopiste tekoälyn sisäisiin ominaisuuksiin. On kehitettävä ymmärrystä siitä, kuinka tekoälyjärjestelmiä voidaan suojata kielteiseltä käytöltä. Tämä tarkastelu on tehtävä aina suurista yrityksistä kahviloissa istuviin yksilöihin. He kaikki ovat alttiita tekoälyyn kohdistuville hyökkäyksille (Kilpatrick, 2019). Tekoälyn varjopuolien kanssa samoilla linjoilla on myös Lehto. Lehto (2019) nostaa esille tekoälymahdollisuuksien lisäksi myös kehitykseen liittyvät turvallisuuskysymykset. Kun päätöksenteko siirtyy ihmiseltä koneelle, mitä meidän on huomioitava? Avointen turvallisuuskysymysten määrää tulee vähentää tekemällä tutkimusta ja antamalla koulutusta (Lehto, 2019). Tämän pro gradun motivaation lähteenä on edellä mainitun varjopuolen tuominen osaksi keskustelua. Tarkoitus on tutkia yhtä tekoälyn turvallisuuskysymyksen osa-aluetta, sen haavoittuvuuksia.

Tekoälyohjelman loppuraportin (2019) ja Tarkoman (2017) mukaan tekoälyn turvallisuuteen liittyvät riskit voidaan luokitella kolmeen eri ryhmään: haitallinen tekoäly, erehtyvä tekoäly sekä tekoälyjärjestelmään kohdistuva haitallinen vaikuttaminen. Haitallinen vaikuttaminen on esimerkiksi hyökkäyksiä tekoälyjärjestelmiä vastaan. Haitallinen tekoäly vaikuttaa ihmisiin tai toisiin järjestelmiin edistääkseen annettua päämäärää. Tekoäly voi myös erehtyä, jos sen kehittämiseen tai päivittämiseen käytetty data on vinoutunutta (Tekoälyohjelman loppuraportti, 2019 ja Tarkoma, 2017).

1.1 Tutkimuksen tausta

> "Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world." Venäjän presidentti Vladimir Putin.

Tämän pro gradu tutkimuksen taustalla on tekoälyn maailmanvalloituksen haasteet. Tekoäly on tekniikka, jossa on potentiaalia niin hyvään kuin pahaan. Bradleyn (2019) mukaan monet merkittävät tieteen ja tekniikan henkilöt kuten Stephen Hawking ja Elon Musk ovat sitä mieltä, että kaikista nykyisistä uhista huipputiede on suurin riski ihmiskunnalle. Se on uhka, joka ylittää huomattavasti ilmastonmuutoksen, ylikansoituksen ja ydinsodan riskit (Bradley, 2019). Myös Floridin ym. (2018) mukaan koko maailma on kohtaamassa tekniikan, joka pitää hallussaan positiivisia lupauksia ihmiskunnan monille osa-alueille. Lisäperusteluja tähän antaa Amodei ym. Amodein ym. (2016) tekemän raportin mukaan järjestelmät, jotka tyypillisesti antavat vain suosituksia ihmisille omaavat suhteellisen rajallisen potentiaalin haitan aiheuttamisessa. Sitä vastoin järjestelmät, jotka hallitsevat isoja kokonaisuuksia voivat aiheuttaa ihmisille rajattomasti haittoja. Näitä haittoja ei välttämättä voi korjata (Amodei ym., 2016). Floridin ym. (2018) ja Amodein ym. (2016) kanssa samaa mieltä on Patel, Hatzakis, Macnish, Ryan ja Kirichenko (2019). Heidän kirjoittamansa raportin mukaan tekoälyä käytetään automatisoitujen päätösten tekemiseen yhä useammissa ympäristöissä. Tämän seurauksena ihmisen osallistuminen päätöksentekoprosesseihin vähenee. Tällöin on luonnollista olettaa, että pahantahtoiset toimijat kiinnostuvat lopulta tekoälyn haavoittuvuuksista. Heidän raporttinsa mukaan tähän löytyy jo esimerkkejä. Muun muassa sosiaalisten verkostojen suositusjärjestelmiä ohjaavat algoritmit ovat joutuneet hyökkäyksen kohteeksi jo vuosien ajan (Patel, Hatzakis, Macnish, Ryan & Kirichenko, 2019). Comiterin (2019) mukaan tämä tarkoittaa sitä, että tekoälyjärjestelmien haavoittuvuudet ovat jo nyt konkreettisia.

Ailiston ym. (2018) mukaan tekoäly on väline, jonka avulla koneet, laitteet, ohjelmat, järjestelmät sekä palvelut voivat toimia tehtävän tai tilanteen mukaisesti järkevällä tavalla. Toiminnan järkevä taso edellyttää, että tekoäly osaa tunnistaa erilaisia tilanteita ja ympäristöjä. Sen on osattava toimia muuttuvien tilanteiden mukaan. Nämä ominaisuudet vaativat tekoälyltä autonomisuutta, oppivuutta ja suorituskykyä. Koska tekoälyn on tunnistettava jatkossa yhä erilaisimpia tilanteita, sen on tavoitetilassa osattava toimia ilman jokaiseen tilanteeseen ennalta tehtyä ohjelmointia. Sen on suoriuduttava sille määritetyistä tehtävistä mielekkäällä tavalla (Ailisto ym., 2018). Ailiston ym. kanssa samaa mieltä on Fralick (2019), jonka mukaan tekoäly on erityisen kyvykäs uhkien metsästyksessä ja havaitsemisessa. On kuitenkin tunnistettava, että tätä se ei voi tehdä ilman omia puutteitaan. Puutteiden takia tekoälyn tietoja tai algoritmeja voidaan manipuloida. Lopputulemana vääristynyt tekoälyjärjestelmä joutuu vaikeuksiin. Tekoälyn toimintaa hankaloittavat myös haittaohjelmat. Haittaohjelmat voivat levitä huomaamatta, mikä vaarantaa tärkeät tiedot, järjestelmät, käyttäjät sekä itse tekoälyjärjestelmät. Tämän takia on kyettävä jatkuvasti seuraamaan

tekoälypohjaisia järjestelmiä. Seuraamalla varmistetaan, että ne tekevät sen mitä niiden on tarkoituskin tehdä. Samalla kyetään jäljittämään niiden kehittymistä ja mukautumista muuttuvaan kontekstiin (Fralick, 2019).

Tekoälyä kohtaan kohdistuvat uhat, eivät poikkea muiden potentiaalisten suorituskykyjen kokemista vaaroista. Brundagen ym. (2018) ja Patelin ym. (2019) mukaan tekoälyn käytön lisääntyessä voimme olettaa, että sen haavoittuvuuksien kautta tapahtuu väärinkäyttöä. Tähän väärinkäyttöön liittyy jo olemassa olevien uhkien laajentaminen, uusien uhkien esittely, uhkien vaihtuminen tai kohdentuminen. Kuten tavanomaisessa kyberuhkassa tai monimutkaisten hyökkäysmenetelmien käytössä, laajat resurssit omaavien toimijoiden kehittämät työkalut päätyvät lopulta rikollisjärjestöjen tai tietoverkkorikollisten käsiin. Sama suuntaus odottaa hyökkäyksiä, jotka on kehitetty tekoälyä vastaan. Väärinkäyttö liittyy suoraan tekoälyn kasvavaan käyttöön ja todennäköisesti niissä kaikissa hyödynnetään tekoälyjärjestelmien haavoittuvuuksia (Brundage ym., 2018 ja Patel ym., 2019).

Myös Suomessa ollaan hereillä uusien teknologioiden haavoittuvuuksien kanssa. Ollilan (2019) mukaan nykyisen teknologiakeskustelun ytimessä on tekoäly, joka on yksi neljännen teollisen vallankumouksen osa-alueista ja keskeinen osa digitalisoituvaa maailmaa. Tekoälyyn liittyvistä uusista keksinnöistä ja aluevaltauksista uutisoidaan päivittäin (Ollila, 2019). Siukosen ja Neittaanmäen kirja tukee myös Ollilan kirjoittamia havaintoja. Siukosen ja Neittaanmäen (2019) mukaan digitalisoituva maailma tuo mukanaan rikollisten mahdollisuuden puuttua tavallisten ihmisten elämään. Tämä voi tapahtua esimerkiksi hakkeroimalla laitteiden toimintaa etäyhteyden kautta. Uusina uhkakuvina on tekoälyjärjestelmien toimintojen sabotointi tai sieppaaminen. Näitä ovat esimerkiksi robottiautot tai droonit (Siukonen & Neittaanmäki, 2019).

Tekoälyn haavoittuvuudet voivat aiheuttaa harmia tai pahimmillaan johtaa onnettomuuksiin. Amodei ym. (2016) mukaan hyvin usein tekoälyonnettomuutta voidaan kuvata tilanteeksi, joka johtui inhimillisen suunnittelijan tavoitteesta tai tehtävästä. Suunnittelijan yllättää se, että järjestelmä tuottikin haitallisia ja täysin odottamattomia tuloksia. Tämä tilanne nousee esimerkiksi suunnitteluvaiheessa (Amodei ym., 2016). Comiter antaa esimerkin suunnitteluvaiheen virheeseen. Comiterin (2019) mukaan itse ajava auto osui kuolettavasti jalankulkijaan Arizonassa. Syyksi paljastui auton sisäinen tekoälyjärjestelmä. Se ei onnistunut havaitsemaan ihmistä. Tämä syy olisi voinut johtua muustakin, kuin suunnitteluvirheestä. Johtopäätöksenä voidaan todeta, että ei ole merkitystä onko onnettomuuden takana inhimillinen virhe vai ilkeä toimija. Suunnitteluvirheet on paikattava ja viat korjattava. Tämä reaalimaailman esimerkki on konkreettinen huomio siihen, että tekoälyn toimintaa tulee tutkia (Comiter, 2019).

1.2 Tutkimuksen tavoite

Tekoälyn on sanottu olevan osa tulevaisuuden jokapäiväisiä ratkaisuja. Sillä on suorituskykyä ja näkyvyyttä, mutta myös uhkakuvia. Patelin ym. (2019) mukaan tekoälypohjaiset järjestelmät yleistyvät, jonka takia myös pahantahtoiset toimijat

haluavat oppivat hyödyntämään niitä. Joitakin reaalimaailman tekoälyjärjestelmiä vastaan on jo hyökätty. Toimet näiden hyökkäysten estämiseksi ovat vielä lapsenkengissä. Tämä on herättänyt tutkijoiden huomion ja todennäköisesti mielenkiinto asiaa kohtaan kasvaa edelleen (Patel ym., 2019).

Patelin ym. kirjoittaman raportin henki näkyy myös tekoälyohjelman loppuraportissa. Tekoälyohjelman loppuraportin (2019) ja Euroopan komission (2019) julkaiseman raportin mukaan tekoälyyn sisältyy valtavasti potentiaalia ja muutosvoimaa. Se voi auttaa meitä ratkaisemaan globaaleja ongelmia, mutta samalla tekoäly luo uusia haasteita yhteiskunnalle. On meistä kiinni, toteutuuko tekoälyn hyötypotentiaali ja vai sen riskit. Luotettava tekoäly edellyttää, että maksimoidaan tekoälyjärjestelmien hyödyt ja ennaltaehkäistään sen tuomat riskit (Tekoälyohjelman loppuraportti, 2019 ja European Comission, 2019). Greimanin (2020) mukaan yksi suurimmista haasteista tutkijoille, päättäjille, hallituksille ja yksityiselle teollisuudelle on se, miten luoda vähemmän keinotekoisia, puolueettomia, älykkäämpiä ja inhimillisempiä tekoälysysteemejä. Jos aiomme edistää tekoälyn järjestelmien maailmanlaajuista kehitystä, on vielä paljon työtä puitteiden kehittämiseksi. Vastuu on jaettava kaikille niille, jotka suunnittelevat, omistavat tai käyttävät älykkäitä järjestelmiä (Greiman, 2020).

Tekoälyn erilaiset vaikutustavat ovat moninaisia. Ollilan (2019) mukaan tekoälyn vaikutus välineenä on poikkeuksellinen. Se muuttaa käyttäjänsä toimintatapaa ja tulee jatkossa olemaan osa käyttäjäänsä. Loppukäyttäjät tulevat monessa yhteydessä samankaltaistumaan tekoälyn kanssa. Oma muuttumisemme tulee olemaan konkreettista. Kun suunnittelemme tekoälyä, suunnittelemme samalla millaisia loppukäyttäjät ja yhteiskunta ovat tulevaisuudessa (Ollila, 2019). Samaa kirjoittaa myös Tarkoma. Tarkoman (2017) mukaan tekoäly on järjestelmä, joka laajamittaisen data-analyysin ja mallintamisen kautta voi vaikuttamaan ihmisiin tai tietojärjestelmiin. Tekoäly voi muuttaa esimerkiksi äänestyskäyttäytymistä kohdentamalla ihmisiin informaatiota (Tarkoma, 2017). Tällöin tekoälyn vaikutus voi kääntyä positiivisesta negatiiviseksi, vaikka toteutetun toimenpiteen takana tätä ei tavoiteltu.

Tähän pohjustukseen liittyen, tässä pro gradu tutkimuksessa tavoitteena on tutkia tekoälyn haavoittuvuuksia. Tarkoitus on tutkia erityisesti niitä haavoittuvuuksia, jotka vaarantavat nykyisten ja tulevien digitaalisten järjestelmien luottamuksellisuuden, eheyden ja saatavuuden kybertoimintaympäristössä. Fyysisessä maailmassa tapahtuvat hyökkäykset voivat olla suunnattu ihmisiin tai infrastruktuuriin. Myös tekoälyn hakkerointi voi olla suunnattu kybertoimintaympäristössä fyysisiin järjestelmiin. Tämä voi aiheuttaa aineellisia vaurioita ja seurauksia. Tekoälyn tietoturvallisuuteen liittyvistä uhkakuvista huolimatta, on muistettava kriittisyys ja puolueeton suhtautuminen erilaisiin väitteisiin. Tekoälymaailmassakaan kaikki ei ole aina niin mustavalkoista mitä kirjoitetaan. Esimerkiksi Ollilan (2019) mukaan tekoälyn uhkien ja mahdollisuuksien liioittelua on meneillään juuri nyt. Osa toiveista ja peloista on selvästi tekoälystä kirjoittavien mielikuvitusta. Ne ovat ennemmin hahmotelmia, kuin olemassa olevaa realismia. Tämä on ymmärrettävää, koska tekoälyn hahmotelmille on saatu ja saadaan jatkossakin lisävirtaa esimerkiksi scifistä. Toisaalta tekoälyn tuomaa autuutta käsittelevät artikkelit ovat usein tuote-esittelyjä, joilla tarjotaan

suurenmoisia mahdollisuuksia myynninedistämistarkoituksessa. Kriitikot taas maalaavat perusuhkia jo virkansa puolesta (Ollila, 2019).

1.3 Tutkimuksen rajaus

Tieteellinen tutkimus tarvitsee rajauksia. Kanasen (2014) mukaan tutkimusaihe on usein liian laaja kokonaisuus. Jotta kokonaisuus saadaan maaliin, tutkimusaihetta pitää rajata. Ilman rajauksia tutkimuksesta tulee pinnallinen, eikä sitä kyetä hallitsemaan perinteisin menetelmin. Tutkimuksen rajaaminen tarkoittaa Kanasen mielestä polun valitsemista. Tällä helpotetaan tutkittavan ilmiön hallintaa. Polun valinta tehdään niin, että tietyt ilmiön osa-alueet otetaan huomioon ja muut jätetään ulkopuolelle. Näin on mahdollista saavuttaa ilmiön hallinta, jolloin tutkimuksella on mahdollisuus päästä kunnialla loppuun saakka. Rajaaminen alkaa usein tutkittavan ilmiön hahmottuessa tutkijalle ja jatkuu tutkimuksen edetessä (Kananen, 2014).

Tässä tutkimuksessa keskitytään tekoälyn haavoittuvuuksiin kybertoimintaympäristössä. Tämän takia tutkimuksesta rajataan pois tekoälyjärjestelmän virheellisestä käytöstä johtuvat haavoittuvuudet. Yleensä virheelliseen käyttöön liittyy inhimillinen tekijä eli loppukäyttäjä. Muun muassa Jääskeläisen (2019) mukaan tekoälyn ongelmat voivat syntyä käyttäjien kautta. Tekoäly oppii esimerkiksi käyttäjien tekemistä valinnoista. Käyttäjien on usein vaikea ymmärtää tätä. Lisäksi liian korkeat odotukset saattavat aiheuttaa loppukäyttäjältä järjestelmän virheellistä käyttöä (Jääskeläinen, 2019). Inhimillisen loppukäyttäjän tekemä virheellinen käyttö on tekoälyn yksi laaja-alaisimmista haavoittuvuuksista. Koska kyseisestä haavoittuvuudesta voisi tehdä oman pro gradunsa useasta eri näkökulmasta, se rajataan pois.

Haavoittuvuuksien tarkastelukulmia ja lähestymistapoja on useita. Tämä pro gradu keskittyy tutkimaan tekoälyn haavoittuvuuksia kybertoimintaympäristön näkökulmasta. Tutkimuksessa ei tarkastella tekoälyn haavoittuvuuksia muusta toimintaympäristönäkökulmasta. Fyysinen, sosiaalinen, henkinen, psykologinen, poliittinen ja sotilaallinen aspekti rajataan pois.

Tekoälyn käyttöön liittyy monia ihmiselämän tunnepohjaisia haavoittuvuuksia. Ollilan (2019) mukaan esimerkiksi tekoälyn hyötydata sisältää ongelmia, ennakkoluuloja, vääristymiä ja syrjintää. Näitä voi kuvata ihmisen elämän moraalisina ongelmina tai haavoittuvuuksina. Nämä ovat asioita, joita emme halua päätyvän tekoälyyn. Usein ihmiset edellyttävät koneiden täydellistä toimintaa, vaikka jollain tavalla niiden mukana on aina erehtyviä ihmisiä (Ollila, 2019). Lehdon (2019) mukaan moni teknologiatutkija näkee nämä tekoälyn eettiset- ja vastuukysymykset isona haasteena. Pohdintaan nousee se, kenellä on eettinen vastuu tekoälyn päätöksenteosta ja kuka vastaa tekoälyn vastuullisesta toiminnasta (Lehto, 2019). Tästä pro gradusta rajataan pois tekoälyn tunnepohjaiset, moraaliset ja eettiset pohdinnat. Tutkimuksen ulkopuolelle jää myös etiikkakysymykset, tekoälyn oikeudenmukaisuuteen sekä ihmisoikeuksiin liittyvät asiat. Vaikka eettistä pohdintaa ei itse tutkimusraportissa ole referoitu, on siihen kiinnitetty

huomiota lähdeainestoa tutkittaessa. Muun muassa Euroopan komission (2019) raporttia: Ethics guidelines for trustworthy AI, on käytetty lähdeaineistona.

Vahvasen (2018) mukaan ihmisen kehittämät koneet voivat nousta tulevaisuudessa ihmisen haastajiksi ja hallitsijoiksi, eli kaikkivaltiaiksi maailman herroiksi. Ne nimetään tieteiskirjallisuudessa supertekoälyksi. Supertekoäly voi olla ristiriidassa ihmisten tahdon kanssa, sekä kehittää oman tahtonsa ja kopioida itseään (Vahvanen, 2018). Tässä pro gradussa ei tarkastella supertekoälyn tai singulariteetin haavoittuvuutta.

Lisäksi tästä pro gradusta rajataan pois tekoälyn yksityisyyden riskit, ihmismäisen hyveen, hyväntahtoisuuden, vastuullisuuden ja valintatilanteiden aiheuttamat tekoälyn haavoittuvuudet. Ollilan (2019) mukaan näitä edellä mainittuja asioita käsitellään usein esimerkillä, jossa tekoälyn valittavana on aikuisen tai lapsen henki. Toinen esimerkki on itseohjautuvien autojen autonominen päätöksenteko vakavaan liikenneonnettomuuteen johtavassa tilanteessa, jossa täytyy tehdä päätös niin sanotusti ”oman” tai ”vastapuolen” välillä. Kolmas esimerkki on irronnut junanvaunu, joka voitaisiin ohjata kahdelle eri raiteelle. Molemmissa vaihtoehdoissa ihmishenkiä tullaan menettämään, mutta mahdollisten uhrien lukumäärä on eri valitusta vaihtoehdosta riippuen (Ollila, 2019).

1.4 Tutkimusongelma ja tutkimuskysymykset

Tieteellinen tutkimus saa inspiraationsa ongelmasta tai ongelmista. Kanasen (2014 ja 2015) mukaan tieteellisessä työssä pitää aina olla ongelma. Tieteellistä tutkimusta ei voida tehdä ilman ratkaistavana olevaa pulmaa eli tutkimusongelmaa. Kun tutkimusongelma on saatu prosessiin, kirjoitetaan se tutkimuskysymyksiksi. Tutkimuskysymyksiin vastaamalla tutkimusongelma tulee ratkaistuksi. Kananen (2014 ja 2015) jatkaa korostamalla tutkimusongelman määrittelyn tärkeyttä. Tutkimusongelma ohjaa koko tutkimusprosessia. Jos tutkimusongelma asetetaan väärin, tutkimuskysymyksetkin ovat vääriä. Tämä johtaa siihen, että menetelmä ja aineisto eivät tuo oikeaa tulosta. Tutkimusongelman ja -kysymysten muotoiluun kannattaa käyttää aikaa, jotta tutkimustyö onnistuu (Kananen, 2014 ja 2015).

Tutkijan oma motivaatio on tärkeä tekijä. Motivaatiota lisää usein tutkimusongelmaa kohtaan oleva mielenkiinto. Laineen, Bambergin ja Jokisen (2007) mukaan tutkijalla on usein hieman aiempaa tietoa tai tietämystä tutkimuksensa kohteena olevasta ilmiöstä. Esimerkiksi tämän aiemman tiedon tai tietämyksen pohjalta saadaan muodostettua tutkimusongelma. Ongelmasta muodostetaan tutkimuskysymyksiä. Prosessi jatkuu, kun kysymys tai kysymykset ohjaavat tutkijan tarvittavan tutkimusaineiston perään (Laine, Bamberg & Jokinen, 2007). Kananen (2014) täydentää, että usein yksi tutkimuskysymys ei riitä. Tällöin tarvitaan tueksi yksi tai useampia apukysymyksiä (Kananen 2014).

Tämän pro gradun tutkimusongelmana on tekoälyn haavoittuvuudet. Ongelmaa on haluttu rajata tarkastelemalla haavoittuvuuksia vain selkeästi määritetyltä alueelta eli kybertoimintaympäristöstä. Tutkimusongelmaksi on haluttu nostaa ajankohtainen ja nyky-yhteiskunnan digitalisoitumiseen liittyvä aihe.

Tekoäly ja sen haavoittuvuudet ovat jo täällä. Muun muassa Halusen raporttien mukaan tekoäly elää nyt keskeistä ajanjaksoa. Halusen (2018b) mukaan nyt ja tulevaisuudessa tekoälyä pyritään hyödyntämään yhä enemmän ja enemmän. Tekoälyä on nykyisin lähes jokaisella elämän- ja elinkeinoalalla. Koska tekoäly tuo uusia mahdollisuuksia monille aloille, luo se myös mahdollisuuksia väärinkäytöksille. Erilaiset pahantahtoiset toimijat voivat pyrkiä toimimaan tekoälyä vastaan, tavoitteena saada se menettelemään omien tarkoitusperiensä mukaisesti (Halunen, 2018b). Yksi vaihtoehto tämän toteuttamiseksi on hyökätä tekoälyä vastaan sen haavoittuvuuksien kautta.

Tässä pro gradussa on määritetty tutkimusongelma ja lähestytty sitä muuntamalla se tutkimuskysymyksiksi. Tutkimuksen päätutkimuskysymys on:

- Mitkä ovat tekoälyn haavoittuvuudet kybertoimintaympäristön näkökulmasta?

Päätutkimuskysymystä tukevat apututkimuskysymykset ovat:

- Minkälaisia hyökkäyksiä tekoälyä vastaan on?
- Miten tekoälyn haavoittuvuudet ja kyberturvallisuus linkittyvät toisiinsa?

1.5 Keskeiset käsitteet

Kanasen (2013) mukaan käsitteiden merkitys tieteessä on erittäin tärkeää. Ne muodostavat tieteellisen toiminnan perustan. Käsitteiden merkitys unohtuu usein, sillä niitä pidetään itsestäänselvyytenä tai sitten niihin ei kiinnitetä huomiota. Keskeiset käsitteet on määriteltävä. Ne kertovat raportin lukijalle, miten kirjoittaja on ymmärtänyt työn keskeisen terminologian. Käsitteiden avulla hallitaan tutkimuksessa esiintyviä ilmiöitä (Kananen, 2013).

Tämän pro gradu tutkimuksen keskeiset käsitteet tekoälystä eivät ole kaikilta osin vakiintuneet. Toisaalta määritelmiä hienosäädetään koko ajan, mikä näkyy määritelmätulvana. Määritelmätulvaan on tässä pro gradussa reagoitu niin, että tietyissä käsitteissä on pitäydytty koko pro gradu tutkimuksen ajan. Määritelmiä on tekoäly-ympäristössä useita ja osa niistä poikkeaa vain hieman toisistaan. Myös kotimaisten ja ulkomaisten terminologioiden, sekä niiden välinen määrittely on keskeistä ymmärtää. Esimerkiksi käsitteiden kääntäminen muun kielisistä teoksista suomen kielelle on selkeästi vaikuttanut Suomessa käytettyyn termiin. Tällöin tulee esille ristiriitaisia tilanteita, joissa sama niin kutsuttu ”suora suomennos” tarkoittaa useampaa eri kontekstia tekoäly-ympäristössä. Käsitteiden kirjavuuteen ovat kiinnittäneet huomiota myös muut. Muun muassa Ollilan (2019) mukaan tekoälykeskustelua vaivaa epämääräinen terminologia, jonka johdosta tekoälyn käsite elää omaa elämäänsä eri käyttöyhteyksissä. Esimerkiksi Kerns (2017) sanoo, että tekoälyä ei ole määritelty konkreettisesti.

Tässä pro gradussa käytetään ensisijaisesti tieto- ja järjestelmäteknisissä lähdeaineistoissa, sekä julkisissa raporteissa käytettyjä yleiskäsitteitä.

Käsitteiden määrittelyyn on päätetty käyttää tuoreiden julkaisujen ammattisanastoa niiden fyysiseen teokseen kirjoitetun arvon ja tuoreuden takia. Tässä tutkimustyössä käytetyt keskeiset käsitteet on avattu lyhyesti ja selkeästi. Tärkeimmät käsitteet tekoäly ja kybertoimintaympäristö on avattu laajemmin. Tekoälykäsitteen määrittelyyn on tietoisesti haettu laaja-alaista näkemystä useilta tekoälytutkijoilta.

Keskeiset käsitteet ovat linkittyneet vahvasti toisiinsa. Tämä tulee ymmärtää myös tekoälyn maailmaan perehdyttäessä. Ailisto ym. (2018) antaa tähän tiivistettyjä esimerkkejä. Heidän kirjoittaman raporttinsa mukaan tekoälyn laskentaympäristöt, alustat ja palvelut liittyvät esimerkiksi pilvipalveluihin. Hahmontunnistus liittyy konenäköön, kuva-analyysiin ja paikannukseen. Luonnollinen kieli ja kognitio liittyy konekääntämiseen, puheentunnistukseen sekä älykkääseen tekstinsyöttöön. Vuorovaikutukseen liittyy esimerkiksi suosittelujärjestelmät, palvelurobotit, chat-botit ja niin sanotut henkilökohtaiset avustajat. Nämä henkilökohtaiset avustajat on rakennettu palvelemaan ihmistä ja toimimaan interaktiossa hänen kanssaan. Ailisto ym. (2018) jatkavat, että järjestelmätasoon ja systeemivaikutuksiin liittyy datapohjaisten ja symbolisten tekoälymenetelmien yhdistäminen. Robotiikka ja koneautomaatio liittyy laitteisiin, jotka kykenevät vaikuttamaan fyysiseen ympäristöönsä esimerkiksi tarttujan, käsivarren, pyörien tai jalkojen avulla. Robotit toimivat usein tietokoneohjelman ohjaamina ja niillä voi olla aisteja esimerkiksi konenäkö sekä tuntoaisti (Ailisto ym., 2018).

1.5.1 Algoritmi

Fryn (2018) sekä Siukosen ja Neittaanmäen (2019) mukaan algoritmi on sarja loogisia ohjeita. Ne kertovat alusta loppuun, miten jokin tehtävä on suoritettava. Algoritmi on yksityiskohtainen kuvaus tai ohje, miten tehtävä tai prosessi suoritetaan. Tarkemmin se koostuu järjestyksessä olevista yksiselitteisistä toiminnoista, jotka voidaan suorittaa ja jotka määrittelevät lopputulokseen johtavan prosessin. Algoritmeille syötetään todellista maailmaa koskevia tietoja eli dataa ja ne ovat melkein aina jonkinlaisia laskutoimituksia. Lopuksi algoritmeille annetaan tavoite ja ne laitetaan töihin toteuttamaan laskutoimituksia annetun tavoitteen saavuttamiseksi. Algoritmeja on lukemattomasti ja useita erilaisia. Jokaisella niistä on oma tavoitteensa, erityispiirteensä sekä hyvät ja huonot puolensa (Fry, 2018 ja Siukonen & Neittaanmäki, 2019).

Kernsin (2017) mukaan tekoäly sisältää edistyneitä algoritmeja, jotka seuraavat matemaattista toimintoa. Siukonen ja Neittaanmäki (2019) jatkavat algoritmin olevan täsmällinen matemaattinen kuvaus tietokonejärjestelmän tai -laitteiston ongelmanratkaisuun tai tehtävän tekemiseksi vaadittavasta toteutuksesta. Se pitää sisällään sääntöjä, käskyjä ja toimintaohjeita. Algoritmi on esimerkiksi valitulla ohjelmointikielellä kirjoitettu tietokoneohjelma. Kun tuhansia algoritmeja yhdistetään järjestelmiksi, saadaan tietokoneista irti erilaisia toimintoja. Tällöin tosin myös epävarmuus lisääntyy (Siukonen & Neittaanmäki, 2019).

1.5.2 Big data

Siukosen ja Neittaanmäen (2019) mukaan big datalla tarkoitetaan massiivisia, jatkuvasti kasvavia, strukturoituja ja ei-strukturoituja tietoja. Big data on tekstiä, äänitteitä, kuvia tai videoita sisältävien tietojoukkojen säilyttämistä, keräämistä ja tiedon käyttämistä. Näiden massiivisten datamäärien hallitseminen ja tietojen analysoiminen on perinteisillä työkaluilla vaikeaa. Tekoäly on yksi ratkaisu big datan koneelliseen käsittelyyn. Big datan avulla ihmiskunta on päässyt käsittelemään, tutkimaan ja ratkomaan ongelmia suurten datamassojen tukemana. Big dataan ei tulisi uskoa sokeasti, sillä informaation omistajilla on valtaa. Valta syntyy tiedon myymisestä, ostamisesta, keräämisestä, tallentamisesta ja jakamisesta. On hyvä ymmärtää, että tietoa voidaan käyttää uudelleen, kopioida ja muuttaa (Siukonen & Neittaanmäki, 2019).

1.5.3 Haavoittuvuus

Laarin toim. (2019) mukaan haavoittuvuus on alttius uhkille. ”Haavoittuvuus voi olla mikä tahansa heikkous, joka mahdollistaa vahingon toteutumisen tai jota voidaan käyttää vahingon aiheuttamisessa. Haavoittuvuuksia voi olla tietojärjestelmissä, prosesseissa ja ihmisen toiminnassa.” Laari toim., 2019, s.29. Laari toim. (2019) jatkaa, että haavoittuvuuksilla tarkoitetaan usein tietojärjestelmien ja ohjelmistojen ei-toivottuja ominaisuuksia. Näiden ei-toivottujen ominaisuuksien vuoksi järjestelmiä voidaan käyttää suunnittelemattomalla tavalla tai väärinkäytön kohteena. Esimerkiksi kybertoimintaympäristössä käytettävät uhkamenetelmät perustuvat järjestelmissä oleviin haavoittuvuuksiin ja niiden hyödyntämiseen. Järjestelmäkokonaisuuksiin voi liittyä myös inhimillisiä haavoittuvuuksia. Inhimillisiä haavoittuvuuksia ovat esimerkiksi puutteelliset tietoturvaohjeistukset, koulutustaso tai virheet prosesseissa (Laari toim., 2019).

1.5.4 Heikko/ kapea ja vahva/ yleinen tekoäly

Jääskeläisen (2019) mukaan tekoäly voidaan jakaa kapeaan ja yleiseen tekoälyyn. Kapea tekoäly on tehty ratkaisemaan jotain tiettyä tai jotain ennalta tarkasti määriteltyä ongelmaa. Yleinen tekoäly pystyy tekemään päätöksiä itsenäisesti ja sillä on ihmisen kaltainen ymmärrys ja tietoisuus. Yleinen tekoäly kykenisi hahmottamaan suuria kokonaisuuksia ja pystyisi tekemään suunnitelmia sekä päätöksiä itsenäisesti. Vuonna 2020 kaikki käytössä oleva tekoäly on kapeaa (Jääskeläinen, 2019).

Siukosen ja Neittaanmäen (2019) mukaan tekoäly voidaan jakaa heikkoon ja vahvaan tekoälyyn. Heikolla tekoälyllä tarkoitetaan yksitäisiin tehtäviin kykeneviä algoritmeja ja koneoppimiseen perustuvia tietokoneohjelmistoja. Ne suoriutuvat tehtävistään algoritmien ansiosta (Siukonen & Neittaanmäki, 2019). Kernsin (2017) mukaan jopa edistyneitä shakkiohjelmia pidetään heikkoina tekoälyinä. Tämä saattaa johtua eroista valvotun ja valvomattoman ohjelmoinnin välillä.

Kernsin (2017) mukaan monissa elokuvissa esillä oleva vahva tekoäly toimii kuten ihmisaivot. Kun kysyt avainsanoilla terästettyjä kysymyksiä, niihin ei ole määritettyjä vastauksia. Vastaus voi jäljitellä avainsanoja, mutta siitä ei voi olla varma. Voit vain olettaa mitä vahva tekoäly vastaa kysymykseesi. Siukosen ja Neittaanmäen (2019) mukaan vahva tekoäly on ihmisviisauden kaltaista tietoisuutta. Se on täysin ihmisestä irrallaan toimivaa älyä. Siinä koneet tai laitteistot oppivat ensin matkimalla ihmisen aivotoimintaa. Tämän jälkeen ne muodostavat sähköistä tietoisuutta. Tällöin kone on osa reaalimaailmaa ja pystyy määrittämään omat pyrkimyksensä ja tavoitteensa (Siukonen & Neittaanmäki, 2019).

Edellä avattujen määritelmien voidaan todeta olevan hyvin lähellä toisiaan. Tässä tutkimuksessa heikko tekoäly on niin lähellä kapeaa tekoälyä, että niitä pidetään yhdenvertaisina. Vahvan tekoälyn määritelmä on niin lähellä yleistä tekoälyä, että niitä pidetään yhdenvertaisina.

1.5.5 Koneoppiminen

Vahvasen (2018) mukaan koneoppimisessa ihmisen ei tarvitse ohjelmoida kaikkia koneen ominaisuuksia tai tietoja. Koneoppimisessa kone oppii itse ympäristöstään ja saavuttaa autonomisesti sille asetettuja päämääriä (Vahvanen, 2018). Siukosen ja Neittaanmäen (2019) mukaan koneoppiminen on tekoälyn osa-alue, jonka tarkoituksena on saada ohjelmisto toimimaan paremmin pohjatiedon ja käyttäjän toiminnan perusteella. Koneoppimistilanteessa kone oppii toistoilla ilman, että sitä erikseen opetetaan. Koneoppimisella pyritään automatisoimaan tiedon tulkintaa ja laajentamaan koneen havainnointikykyä. Tämä tiedon tulkinta ja koneen havainnointi tapahtuu monimutkaisten algoritmien avulla (Siukonen & Neittaanmäki, 2019). Ailiston, ym. (2018) mukaan koneoppimisen menetelmät liittyvät vahvistettuun oppimiseen, ohjattuun ja ohjaamattomaan oppimiseen. Nämä liittyvät esimerkiksi kasvojentunnistukseen, kuvahakuihin, ja autonomisiin ajoneuvoihin, jossa on erityisesti konenäköön perustuvaa ohjausta.

1.5.6 Kybertoimintaympäristö

”Kybertoimintaympäristö on digitaalisen informaation käsittelyyn tarkoitettu, toisiinsa yhteydessä olevista tietokoneista ja muista laitteista sekä tietoverkoista muodostunut ympäristö.” Lönnqvist & Moilanen, 2017, s.7.

Laarin toim. (2019) mukaan kybertoimintaympäristö on digitaalisista tietojärjestelmistä muodostuva toimintaympäristö. Siihen kuuluvat fyysiset rakenteet, sekä kaikki toimintaympäristön toimijat. Kybertoimintaympäristölle on tunnusomaista elektroniikan ja sähkömagneettisen spektrin käyttö. Lisäksi siihen kuuluu datan sekä informaation varastointi, muokkaaminen ja siirto viestintäverkkojen avulla. Kybertoimintaympäristö ei ole maantieteellisesti rajoitettu ja sen etäisyyttä tarkastellaan eri tavoin kuin perinteistä toimintaympäristöä. Asia tulee parhaiten esille siinä, että kybertoimintaympäristön komponentti saattaa sijaita fyysisesti toisella puolella maailmaa kuin sen loppukäyttäjä. Sijaintiriippumattomuus aiheuttaa sen, että maailmanlaajuinen ympäristö on haavoittuva useista eri kohdista. Tämä johtuu siitä, että kybertoimintaympäristökokonaisuus on

rakennettu laajan verkon ympärille, johon on pääsy lähes kaikkialta. Kybertoimintaympäristöä ei omista kukaan tai toisaalta sen omistavat kaikki sitä käyttävät. Sitä käyttää liike-elämä, valtiot ja yksilöt (Laari toim., 2019).

Laarin toim. (2019) mukaan internetillä on keskeinen asema kybertoimintaympäristössä. Internet on kybertoimintaympäristön yhdistävä tekijä, mutta ei ainut toimija. Kybertoimintaympäristö sisältää esimerkiksi teollisuusautomaatiota, ohjausjärjestelmiä, toiminnanohjausjärjestelmiä, esineiden internetin sekä internettiin liitettyjä tai täysin irrallisia tietoverkkoja. Kybertoimintaympäristöä hahmotettaessa on keskeistä tunnistaa mitä päivittäisessä käytössä ei huomaa. Esimerkiksi laaja osa yhteiskunnan elintärkeistä toiminnoista ja kriittisestä infrastruktuurista on verkottunut, vaikka sitä ei tunnista jokapäiväisessä elämässä (Laari toim., 2019).

Laarin toim. (2019) mukaan kybertoimintaympäristö on lähes kaikkien ulottuvilla. Yleensä kybertoimintaympäristöön liitytään tabletilla, pöytätietokoneella, kannettavalla tietokoneella tai matkapuhelimella. Yhteys saavutetaan langattomien yhteyksien tai fyysisten kuparikaapeleiden tai valokuitujen avulla. Kybertoimintaympäristö on riippuvainen aina fyysisistä ympäristötekijöistä kuten virtalähteistä, kaapeleista, kuiduista, verkoista ja datakeskuksista. Nämä fyysiset tekijät ja kybertoimintaympäristön muodostavat tekniikat ja järjestelmät ovat kehittyneet nykyaikaisen elämäntavan peruspilareiksi. Nyky-yhteiskunnan perustoiminnot ovat riippuvaisia tietovirroista, joten kybertoimintaympäristö on olennainen osa nykypäivän globaalia toimintaympäristöä (Laari toim., 2019). Keskeisessä asemassa on kuitenkin kybertoimintaympäristön järjestelmien, laitteiden ja fyysisten tekijöiden loppukäyttäjä eli ihminen.

1.5.7 Kyberturvallisuus

Järvisen (2018) mukaan kreikkalaiset keksivät kyberin. Kreikan sana cybernetice (tai kubernetes) tarkoittaa ohjausta ja hallintaa. Nykyään kyber-alkuisia sanoja tavataan monissa eri yhteyksissä antamassa termeille lisää dramaattisuutta ja ajankohtaisuutta. Kyberturvallisuus kuulostaa tekniseltä, mutta se on maanläheistä yhteiskunnan arkipäiväisten järjestelmien suojaamista ja niiden toiminnan turvaamista (Järvinen, 2018).

Järvisen (2018) mukaan kyberturvallisuus itsessään sisältää tietoturvan. Samalla kun huolehdimme tietoturvasta, huolehdimme kyberturvallisuudesta. Tietoturvaan liittyy uhkakuvia tiedostojen tai salasanojen katoamisesta aina laitteiden varastamiseen tai verkkomurtoihin. Kyberuhkakuviin liittyy laajempia ja vaikutukseltaan suurempia kokonaisuuksia (Järvinen, 2018).

Laarin toim. (2019) mukaan kyberturvallisuus on tavoitetila. Tavoitetilassa kybertoimintaympäristöön voidaan luottaa ja jossa toiminta turvataan. Kyberturvallisuudella turvataan tiedon, laitteistojen, verkostojen, ohjelmistojen ja käyttäjien luottamuksellisuus, eheys sekä saatavuus koko elinjakson ajan. Kyberturvallisuus muodostuu ylläpitäjien ja käyttäjien välisestä yhteistoiminnasta. Kyberturvallisuudessa huomioidaan kybertoimintaympäristön vaikutukset fyysiseen maailmaan. Fyysisessä maailmassa pelkkä vahinko tai huolimattomuus voi

vaarantaa esimerkiksi koko verkkopankin, veden- tai sähkönjakelun. Tällöin kärsiviä asiakkaita on paljon enemmän (Laari toim., 2019).

1.5.8 Loppukäyttäjä

Tähtisen (2005) mukaan loppukäyttäjä on henkilö, jonka työkaluna jokin sovellus toimii ja jonka työtehtäviä sovelluksen tulisi helpottaa. Sovelluksilla on usein yksi tai useampi loppukäyttäjä. Loppukäyttäjä palvelee organisaation tavoitteita yhtä tai useampaa sovellusta käyttäen (Tähtinen, 2005).

1.5.9 Singulariteetti

Jääskeläisen (2019) ja Järvisen (2018) mukaan singulariteetti tarkoittaa ihmistä älykkäämmän supertekoälyn syntyä. Nyt käytössä oleva tekoäly on kapeaa tekoälyä. Kapea tekoäly on tehty ratkaisemaan ennalta määriteltyä ongelmaa, mutta siitä edetään kohti yleistä tekoälyä. Yleinen tekoäly pystyy tekemään päätöksiä ihmisen kaltaisesti, hahmottamaan suuria kokonaisuuksia ja tekemään suunnitelmia itsenäisesti. Yleinen tekoäly ei tule jäämään tekoälyalan viimeiseksi vaiheeksi. Tulevaisuudessa riittävän edistynyt yleinen tekoäly voi kehittää itseään supertekoälyksi. Lopulta ihmiset eivät pysy tämän kehityksen mukana. Itseään kehittävä tekoäly johtaa exponentiaaliseen kierteeseen, jossa tekoäly syrjäyttää ihmisen. Tekniikka karkaa tasolle, jota kutsutaan singulariteetiksi tai supertekoälyksi (Jääskeläinen, 2019 ja Järvinen, 2018).

1.5.10 Tekoäly

> Tekoäly tarkoittaa laitteita, ohjelmistoja ja järjestelmiä, jotka kykenevät oppimaan ja tekemään päätöksiä lähes samalla tavalla kuin ihmiset. Tekoälyn avulla koneet, laitteet, ohjelmat, järjestelmät ja palvelut voivat toimia tehtävän ja tilanteen mukaisesti järkevällä tavalla. Tekoälyohjelman loppuraportti, 2019, s. 16.

Hurleyn ja Potterin (2020) mukaan tekoäly on termi, jonka määritelmästä on vaikea rakentaa konsensusta. Tästä huolimatta sitä yritetään määritellä useissa lähteissä. Esimerkiksi Townsendin määritelmän konsensusta on haastava ymmärtää yksiselitteisesti. Townsendin (2018) mukaan tekoäly on tietokoneiden käyttöä sellaisten analyyttisten toimintojen suorittamiseen, jotka ovat normaalisti käytettävissä vain ihmisille mutta koneen nopeudella.

Tekoälyn kokonaiskuva ja osaamiskartoitus -raportissa tekoäly ymmärretään seuraavasti: ”Tekoälyn avulla koneet, laitteet, ohjelmat, järjestelmät ja palvelut voivat toimia tehtävän ja tilanteen mukaisesti järkevällä tavalla.” (Ailisto ym., 2018, s. 1).

Honkelan (2017) mukaan tekoälyssä on kyse siitä, että koneen avulla mallinnetaan tai matkitaan ihmisen älykkääksi katsomiaan toimintoja. Esimerkiksi liikkuminen, vaikuttaminen liikkumiseen, kielenkäyttö, ongelmanratkaisu ja aistijärjestelmien käyttö. Tekoäly ei ole pelkkää tietotekniikkaa tai tietojenkäsittelyä vaan monialainen kokonaisuus (Honkela, 2017).

Jääskeläisen (2019) mukaan tekoälyssä tietokoneet pystyvät toimintaan tilanteissa, joissa on perinteisesti ajateltu vaadittavan ihmisälyä. Tekoälyssä tietokoneet kykenevät itsenäisesti mukauttamaan toimintaansa niille annetun datan perusteella. Tekoälyn keskeisiä sovelluksia on ennustaminen datan perusteella, datan luokittelu, kääntäminen, puheen tunnistaminen, robotiikka ja autonomia (Jääskeläinen, 2019).

Järvisen (2018) mukaan tämän päivän tekoälyllä tarkoitetaan esimerkiksi tietokoneohjelmaa, joka pystyy suorittamaan tehtäviä, joihin aiemmin tarvittiin suorittajaksi ihminen. Tällaisia ovat esimerkiksi tekstin kääntäminen kielestä toiseen, puheen ymmärtäminen, lautapelin pelaaminen ja kuvien tulkinta (Järvinen, 2018). Brundage ym. (2018) täydentää edellistä tiivistäen tekoälyn tarkoittavan digitaalisia ja teknisiä järjestelmiä, jotka kykenevät suorittamaan yleisesti tehtäviä, joihin vaaditaan älykkyyttä.

Siukosen ja Neittaanmäen (2019) mukaan tekoäly tarkoittaa laitteita, ohjelmistoja, palveluita ja järjestelmiä, jotka kykenevät oppimaan sekä tekemään päätöksiä lähes samalla tavalla kuin ihmiset. Tekoälyn avulla laitteet, ohjelmistot, palvelut ja järjestelmät voivat toimia tehtävän sekä tilanteen edellyttämällä järkevällä tavalla. Tekoäly oppii, jonka jälkeen sitä voidaan käyttää muualla hyödyksi reagoimaan erilaisiin ärsykkeisiin sopivalla tavalla. Siukonen ja Neittaanmäki (2019) jatkavat tekoälyn olevan tietokoneen toimintojen jatkeena toimiva ohjelma, ohjelmisto tai järjestelmä, joka kykenee mittaviin laskentoihin. Tekoäly viittaa siis tietokoneen toimintoihin, joihin normaalisti tarvitaan ihmisälyä. Toisaalta tekoäly ilmentyy prosesseissa, joihin ihmistä ei kannata käyttää. Näitä ovat esimerkiksi robotiikka ja automaatio. Mittavien laskentakykyjen takia tekoäly on tietotekniikan, tietojenkäsittelytieteen ja informaatioteknologian osa-alue. Nyt tätä osa-aluetta tutkitaan, jalostetaan, kaupallistetaan, kehutaan ja pelätään. Tekoälyn englanninkielinen nimi artificial intelligence (AI) tarkoittaa tietokoneen tuottamaa keinotekoista älykkyyttä luonnollisen älykkyyden ja oppimisen yhteydessä (Siukonen & Neittaanmäki, 2019).

Ailiston ym. (2018) ja Siukosen ja Neittaanmäen (2019) mukaan tekoäly syntyy useita tietotekniikan osa-alueita yhdistämällä ihmisen ja koneen vuorovaikutuksessa. Tekoälyteknologian alle kuuluu joukko erilaisia menetelmiä, teknologioita ja sovelluksia. Tekoäly on menetelmien, teknologioiden ja sovellusten yksi kehitysaskel digitalisaation laajemmassa viitekehyksessä. Tekoälyjärjestelmän on siis osattava käyttää koneoppimista, syväoppimista, algoritmeja, neuroverkkoja, suuria datamassoja, perinteistä logiikkaa ja sumeata logiikkaa, simulointia, optimointia, mallinnusta, signaalin ja datan käsittelyä, tiedonlouhintaa, konenäköä, konetietoisuutta, puheentunnistusta ja -tuottamista, automaatiota, robotiikkaa sekä monimutkaisten järjestelmien hallintaa ja päätöksentekoa. Tekoäly hyödyntää näitä osa-alueita ja valjastaa ne käyttöönsä. Tekoälyn kehittäjien tahto on saada tietokoneen ohjelmistot oppimaan ihmismäisesti toistojen, erehdyksen ja opetuksen kautta. Tällä hetkellä tekoälylaitteet oppivat ihmisen antamien syötteiden tai ”eväiden” ja ohjelmoijien kirjoittamien algoritmien mukaan (Ailiston ym., 2018 ja Siukonen & Neittaanmäki, 2019). Näitä syötteitä ja ”eväitä” on nykyisin valtavasti. Ne painottuvat tekoälyn eri osa-alueiden mukaan. Esimerkiksi Castelluccion (2018b) mukaan suuria ponnisteluita liittyy neuroverkoihin, jotka laskevat massiivisia tietovarastoja ja oppivat näistä tiedoista. Tämän

takia tekoälyn kehittyminen on sanottu johtuvan tietokoneiden laskentatehon kasvusta. Vahvasen (2018) mukaan kyse ei ole vain laskentatehosta, vaan myös siitä miten sitä käytetään.

1.5.11 Tietojärjestelmä ja tietojärjestelmäkokonaisuus

Tähtisen (2005) mukaan tietojärjestelmä on looginen kokonaisuus, joka ottaa vastaan tietoa, käsittelee sen ja tuottaa jatkotietoa. Tällainen kokonaisuus on esimerkiksi ohjelmistotuote, joka voi olla paketoitu tai räätälöity ohjelmistokokonaisuus. Tietojärjestelmä on yleensä yrityksen sisäisessä käytössä ja hallinnassa. Siitä voidaan lukea informaatiota ja siihen voidaan kirjoittaa informaatiota useilla eri tavoilla (Tähtinen, 2005). Mikkonen (2003) täydentää, että tietojärjestelmistä koostuvaan tietojärjestelmäkokonaisuuteen kuuluvat erilaiset tietoa tuottavat sensorit, anturit, mittarit ja tietoja välittävät tiedonsiirtoverkot. Siihen kuuluu myös alustapalvelut, tietoja kokoavat ja käsittelevät tietojärjestelmät sekä ohjelmat tai palvelut (Mikkonen, 2003).

1.5.12 Vääristymä

Euroopan komission (2019) raportin mukaan vääristymä on puolueellisuutta jotakin kohdetta, henkilöä tai kantaa vastaan. Se voi olla myös sen puolesta. Vääristymiä voi syntyä tekoälyjärjestelmissä monin tavoin. Datavetoisissa tekoälyjärjestelmissä vääristymä saattaa johtaa siihen, että järjestelmä toimii puolueellisesti. Logiikkaan perustuvissa tekoälyjärjestelmissä vääristymä voi tulla siitä, miten tekijä suhtautuu tietyssä tilanteessa sovellettaviin sääntöihin. Vääristymä voi syntyä esimerkiksi oppimisen ja vuorovaikutuksen avulla tapahtuvan mukautumisen seurauksena. Se voi olla yksilöllistä, jos esimerkiksi tekoälyjärjestelmä on räätälöity käyttäjän mukaan. Vääristymä voi olla myönteinen, kielteinen, tahallinen tai tahaton ja voi johtaa syrjiviin ja/tai epäoikeudenmukaisin tuloksiin (European Comission, 2019).

1.6 Katsaus aikaisempiin tutkimuksiin ja kirjallisuuskatsaus

Tarkoman (2017) mukaan tekoäly on monitieteinen tieteenala, joka analysoi älykkääksi katsottua toimintaa ja tutkii älykkäiden järjestelmien tuottamista. Tekoäly pitää sisällään tietojenkäsittelytieteen, matematiikan, tilastotieteen sekä kognitiotieteen menetelmiä. Nämä yhdessä muodostavat älykkään järjestelmän (Tarkoma, 2017). Tarkoman kanssa samoilla linjoilla on Honkela. Honkelan (2017) mukaan tekoälytutkimus on kokonaisuutena hyvin läheinen kognitiotieteiden kanssa. Kognitiotiede on ihmisen ajattelun ja älykkäiden toimintojen tutkimusta. Kognitiotieteen tavoitteet ovat samankaltaisia, kuin tekoälyn kehittämisessä (Honkela, 2017).

Monitieteellisenä tieteenalana tekoäly pitää sisällään positiivisia asioita sekä paljon mahdollisuuksia. Tekoälygenren piirissä on pitkän linjan tutkijoita ja

tieteenalana se kiinnostaa myös tulevaisuuden tutkijoita. Konttisen (2018) mukaan tekoälytutkijat ovat kaikki samaa mieltä siitä, että tekoäly mahdollistaa paljon myönteistä. Samaan hengenvetoon tutkijat toteavat, että tekoälyn väärinkäytöstä on ollut tieteellistä tutkimusta vain nimeksi. Tekoälyn väärinkäytön tutkimisessa ei ole mistään kaukaiseen tulevaisuuteen ulottuvasta näkymästä. Väärinkäyttö on totta jo 2020-luvun alussa. Pahantahtoiset toimijat voivat hyödyntää iskuissaan tekoälyä laajalla rintamalla, täsmällisesti, tehokkaasti ja edullisesti. Tämän ovat ymmärtäneet myös tekoälytutkijat. He ovat huolissaan tekoälyn antamista mahdollisuuksista rikollisille. Tutkijat varoittavat, että tekoälyteknologian haavoittuvuudet voivat mahdollistaa tulevaisuudessa esimerkiksi terroriiskuja. Iskuja voisi toteuttaa esimerkiksi hakkeroimalla tai kaappaamalla tekoälyjärjestelmiä. Käyttäjät luottavat liikaa tekoälyteknologiaan huolimatta siitä, että sekin on murrettavissa ja manipuloitavissa (Konttinen, 2018).

Tämän tekoälytutkimuksen kirjallisuuskatsauksen aikana tuli nopeasti esille, että tutkijoiden mukaan tekoälyllä on useita erilaisia haavoittuvuuksia. Bradleyn (2019) mukaan tekoälyn aiheuttamista riskeistä on olemassa kirjoja, artikkeleita, blogeja, analyyseja ja mielipiteitä. Julkaistua tutkimusta riskien hallitsemisesta on kuitenkin vähän. Vielä vähemmän on käytännön ohjeita työkalujen tai standardien muodossa (Bradley, 2019). Tekoälyn kybertoimintaympäristön haavoittuvuuksista ei ole saatavilla tutkimukseen perustuvaa kirjallisuutta. Brundage ym. (2018) mukaan tekoälyllä on haavoittuvuuksia, heikkouksia ja väärinkäytön mahdollisuuksia useassa eri kontekstissa ja ympäristössä. Koska tekoälyn haavoittuvuuksien konteksti on laaja ja ympäristöjä useita, tässä pro gradu tutkimuksessa keskitytään tekoälyn haavoittuvuuksiin kybertoimintaympäristön näkökulmasta. Kybertoimintaympäristön näkökulmaan liittyy erityisesti teknologinen ja digitaalinen turvallisuusympäristö. Brundagen ym. (2018) mukaan tekoälyn käyttö automatisoiduissa tehtävissä nostaa siihen kohdistuvien kybertoimintaympäristössä toteutettavien hyökkäysten uhkaa.

Tekoälyn väärinkäytöstä on löydettävissä vain vähän tieteellistä tutkimusta, toteaa Konttinen (2018). Tämä sama haaste koskee tekoälyn haavoittuvuuksia. Erityisesti tekoälyn haavoittuvuuksien tutkimusta tai siihen keskittyvää kirjallisuutta ei ole kirjallisuuskatsauksen mukaan saatavilla. Kirjallisuuskatsauksen perusteella haavoittuvuuksien tutkimustulokset tai ilmiöt nousevat esille muiden havaintojen lomassa. Tekoälyn haavoittuvuuksien tutkimus on jäänyt tekoälyn mahdollisuuksien, tekoälypohjaisen puolustuksen ja innovaatiotutkimuksen varjoon. Tekoälyn käyttöä kyberturvallisuudessa on myös tutkittu ja toiminta siinä on melko aktiivista. Edellä mainituissa tutkimuksissa löytyy myös mainintoja tekoälyn haavoittuvuuksista. Tekoälyn haavoittuvuuksien tutkimus on siis jäänyt tekoälyn kokonaisuuden tutkimuksessa vain yhdeksi pieneksi osa-alueeksi. Tämä osa-alue ei ole ilmiselvästi painopisteenä.

Usean kiven kääntämisen jälkeen tutkimusaiheeseen liittyvää aineistoa kuitenkin löytyi. Tosin sitä löytyi niukasti. Lähimpänä tämän pro gradun tutkimusaihetta on EU: n rahoittama Horizon 2020 -hankkeen (koodinimi SHERPA) raportti vuodelta 2019. Horizon 2020 -hankkeen tavoitteena oli lisätä ymmärrystä siitä, miten tekoälyä ja siihen liittyviä osa-alueita käytetään tulevaisuudessa yhteiskunnassa. Tehtävän tulos oli raportti: Security Issues, Dangers and Implications of Smart Information Systems. Raportin kirjoittajina ovat toimineet Patel,

Hatzakis, Macnish, Ryan ja Kirichenko. Tämä lähde on keskeinen tekoälyn haavoittuvuuksien ymmärtämisessä.

Toinen tätä tutkimusta lähellä on useamman instituution ja yliopiston 26 asiantuntijan laatima raportti vuodelta 2018: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Tässä raportissa painopiste on tarkastella tilanteita, joissa henkilö tai organisaatio käyttää tekoälyä. Raportissa pohditaan vaarantaako tekoälyä käyttävä yksilö toisen henkilön tai organisaation turvallisuutta. Raportissa käsitellään monenlaisia turvallisuusuhkia, kuten digitaalista, fyysistä ja poliittista turvallisuutta. Julkaisun keskeinen teema on välitön ja kriittinen tarve kehittää vastuullisuuskulttuuria tekoälytutkimuksessa. Raportti sopii laajempaan tekoälykokonaisuuteen muun muassa tekoälyn sosiaalisista vaikutuksista aina poliittisiin vastauksiin saakka. Tätä raporttia on siteerattu monessa lehtiartikkelissa ja julkaisussa. Muun muassa Castelluccio (2018a) kirjoittaa siitä Strategic Finance -julkaisussa, Roberts (2018) Cambridgen yliopiston julkaisussa ja Bradley (2019) AI & Society -tutkimuslehdessä.

Kolmas tätä pro gradua lähellä oleva raportti on Comiterin kirjoittama ja Harvard Kennedy Schoolin vuonna 2019 julkaisema raportti: Attacking Artificial Intelligence. AI's Security Vulnerability and What Policymakers Can Do About It. Kyseinen raportti painottuu tekoälyä vastaan kohdistuviin hyökkäyksiin ja vain vähän sen haavoittuvuuksiin. Tämä lähde on keskeinen tekoälyä kohtaan tehtävien hyökkäyksien ymmärtämisessä.

Ulkomaisesta kirjallisuudesta tekoälyn haavoittuvuuksista löytyy julkaistua tietoa Fryn (2018): Hello world ja Sautoyn (2019): The creativity code. How AI is learning to write, paint and think. Tekoälyn haavoittuvuuksiin liittyvien ulkomaalaisten kirjojen saatavuudessa on haasteita. Usein ongelmana on julkaistujen kirjojen saatavuus kirjastoissa tai muissa instituutioissa. Asia on toki hieman parantunut e-kirjojen tulon myötä, mutta maksuttomien julkaisujen saatavuudessa näkyy aina kysynnän ja tarjonnan laki.

Julkaistusta suomalaisesta kirjallisuudesta lähimpänä tätä pro gradu tutkimuksen aihetta on Lehdon (2019) kirjoitus: Onko tekoäly turvallinen? Kirjotus on osa Siukosen ja Neittaanmäen (2019) kirjaa: Mitä tulisi tietää tekoälystä, vuodelta 2019. Lisäksi Järvisen: Kyberuhkia ja somesotaa (2018) ja Honkelan: Rauhankone. Tekoälytutkijan testamentti (2017) -kirjat sisältävät tietoa tekoälyn haavoittuvuuksista.

Ulkomaankielisen verkkomateriaalin osalta tekoälyn haavoittuvuuksia on käsitellyt Zheng (2017) tekstissään: The Cybersecurity Vulnerabilities to Artificial Intelligence. Toinen mainitsemisen arvoinen lähde MC.AI kautta löytyvä artikkeli (2019): 9 Critical AI Weaknesses to Consider. Suomenkielisen verkkomateriaalin osalta tämän pro gradun aihetta on käsitellyt VTT:n tutkija Kimmo Halunen. Julkaistuja kirjoituksia tutkimusaiheen osalta on käyty läpi eri näkökulmista VTT:n blogissa. VTT:n blogista nousee keskeisesti esille seuraavat kolme julkaisua: Tekoälykin voi haavoittua eikä täydellistä tekoälysovellusta ole (2018), Miten tekoälyjä harhautetaan? (2019) ja Lohkoketjusta tekoälyn luottamuksen rakentaja? (2018). Erityisesti kaksi ensimmäiseksi mainittua julkaisua tuovat esille tekoälyn haavoittuvuudet ja hyökkäysvektorit. Julkaisut tuovat esille, että tekoälyä voidaan hämätä ja sitä vastaan voidaan hyökätä sen haavoittuvuuksien

kautta. Saadun tiedon mukaan aihealueen syvempi tutkimus ei ole jatkunut VTT:ssä rahoituksellisista syistä.

2 TUTKIMUKSEN TIETEELLINEN POHJA

Tämän pro gradu tutkimuksen tieteellisenä pohjana toimivat tutkimustyyppi, tutkimussuuntaus, tutkimuslaji, aineiston keruumenetelmä, aineiston analyysimenetelmä sekä tutkimuksen luotettavuus.

Tutkimukset jakautuvat pääsääntöisesti laadulliseen, määrälliseen tai näiden yhdistelmää käyttäviin tutkimuksiin. Kanasen (2013) mukaan laadullinen tutkimus on kaiken tutkimustoiminnan perusta. Laadullista tutkimusta voidaan pitää ”tutkimuksen äitinä”, koska myös määrällinen tutkimus perustuu laadulliseen tutkimukseen. Laadullinen tutkimus pyrkii ymmärtämään ilmiötä. Lisäksi se pyrkii selittämään ilmiön koostumusta, tekijöitä ja niiden välisiä suhteita. Laadullinen tutkimus tuottaa selityksen käytännöstä ja vastaa kysymykseen: ”Mistä tässä on kyse?” (Kananen, 2013).

Hirsijärven, Remeksen ja Sajavaaran (2000 ja 2004) sekä Tuomen ja Sarajärven (2000) mukaan tarvitsemme kvalitatiivista eli laadullista tutkimussuuntausta, kun olemme kiinnostuneita asioista, joita ei voi yksinkertaisella tavalla mitata määrällisesti. Tuomen ja Sarajärven (2009) mukaan laadullinen tutkimus on terminä eräänlainen sateenvarjo, jonka alla on useita hyvin erilaatuisia laadullisia tutkimuksia. Laadullisesta tutkimuksesta voidaan puhua sekä laajassa merkityksessä, että kapeassa merkityksessä (Tuomi & Sarajärvi, 2009).

2.1 Tutkimusmetodi ja metodologia

Metodologian ja metodin ero on välillä haastava tunnistaa. Tuomen ja Sarajärven (2009) ja Tuomen (2007) mukaan tutkimuksen aineiston keruu- ja analyysimetodit ovat tutkimustulosten ja sitä kautta syntyneen tiedon perustelu sekä oikeutus. Metodi on selitys sille, miksi tuollaisia tai tällaisia tietoja on tutkimuksessa saatu selville. Metodi perustelee tutkimuksessa syntyneen tiedon ja metodologia kysyy, onko käytetty metodi järkevä. Metodologia on sääntöjä siitä, miten välineitä eli metodeja käytetään (Tuomi & Sarajärvi, 2009 ja Tuomi, 2007). Tuomi (2007) tarkentaa vielä, että laajassa merkityksessä metodologia käsittelee todellisuutta koskevan tiedon peruslähtökohtaa, perusnäkemystä ja maailmankatsomusta. Suppeassa merkityksessä metodologialla tarkoitetaan metodien käyttöä, eli sitä miten tutkimuskäytössä hankitaan uutta tietoa todellisuudesta (Tuomi, 2007). Kananen (2019) kirjoittaa, että metodologia tarkoittaa samaa kuin tutkimusote tai lähestymistapa. Laineen ym. (2007) mukaan tutkijan on mietittävä, millä keinoin lähdeaineisto auttaa vastaamaan tutkimuskysymykseen. Metodologien käyttö on mietittävä suhteessa aineistoon ja aineisto on kerättävä tutkimuskysymys mielessä pitäen (Laine ym. 2007).

Hirsijärven ym. (2000) mukaan metodin eli tutkimusmenetelmän/ tutkimusotteen käsite on moniselitteinen. Metodi koostuu niistä tavoista ja käytännöistä, joilla tutkimuksen havaintoja kerätään. Metodi on sääntöjen ohjaama menettelytapa, jonka avulla tieteessä tavoitellaan ja etsitään tietoa tai pyritään ratkaisemaan käytännön elämän haasteita (Hirsijärvi ym., 2000).

2.2 Tutkimustyyppi

Tämän pro gradun tutkimustyyppi on teoreettinen tutkimus. Tuomen (2007) mukaan teoreettiskäsitteellinen tutkimus edellyttää perehtymistä kirjalliseen aineistoon, jossa argumentaatio muodostaa metodin ydinosan. Teoreettisessa tutkimuksessa ei ole empiiristä havaintoaineistoa ja se ei käytä metodeja argumentoinnin välineenä. Toisaalta tutkimus voisi olla myös empiirinen, koska teoreettisella ja empiirisellä tutkimustyypillä voi tutkia samaa ilmiötä. Tutkimuksellinen ero näiden kahden tutkimustyypin välillä voidaan pelkistää siihen, mistä näkökulmasta havaintoaineistoa tarkastellaan (Tuomi, 2007).

Teoreettisen tutkimustyypin valintaa tukee tässä pro gradu tutkimuksessa käytetty tutkimuslaji, sekä aineiston keräämis- ja analyysimenetelmät. Tutkimuksen aineistonkeruumenetelmä perustuu erilaisiin dokumentteihin. Tämä mahdollistaa aineiston analysoinnin tapahtuvan dokumenttien antamien tietojen ehdoilla, eikä tutkijan ennakkoluulojen mukaan.

2.3 Tutkimussuuntaus

Hirsijärven ym. (2000 ja 2004) sekä Tuomen ja Sarajärven (2000) mukaan, lähtökohtana laadullisessa tutkimuksessa on todellisen kuvaaminen ymmärrettävästi. Todellisuus kuvataan usein moninaisena. Laadullisen tutkimuksen tavoitteena on kuvata kohdetta mahdollisimman kokonaisvaltaisesti ja eläytyä tutkimuskohteeseen. Laadullisessa tutkimuksessa on pyrkimyksenä löytää tai paljastaa asioita. Laadullisessa tutkimuksessa ei ole tarkoitus pyrkiä todistamaan jo olemassa olevia totuusväittämiä (Hirsijärvi ym., 2000 ja 2004, Tuomi & Sarajärvi, 2000).

Hirsijärven ym. (2004) mukaan laadullinen tutkimus on luonteeltaan kokonaisvaltaista. Se on tiedon hankintaa, jonka aineisto kootaan luonnollisissa ja todellisissa tilanteissa. Laadullisessa tutkimuksessa käytetään induktiivista analyysia, jossa tutkijan pyrkimyksenä on paljastaa odottamattomia seikkoja. Tämän takia aineiston monitahoinen ja yksityiskohtainen tarkastelu toimii lähtökohtana. Tutkija ei määrää sitä, mikä on tärkeää. Laadullisen tutkimuksen aineistonhankinnassa käytetään laadullisia metodeja. Niissä tutkittavien näkökulmat pääsevät esille (Hirsijärvi, Remes & Sajavaara, 2004).

Alasuutarin (2001) mukaan laadullisessa tutkimuksessa teoreettinen viitekehys määrää sen, millainen aineisto kannattaa kerätä ja millaista menetelmää sen analyysissä kannattaa käyttää. Teoreettisen viitekehyksen ja sen kanssa sopusoinnussa olevan tutkimusmenetelmän valitseminen on tärkeää. Laadulliselle tutkimukselle on luonteenomaista kerätä aineistoa, joka tekee mahdollisimman monenlaiset aineistotarkastelut mahdolliseksi. Laadullisen tutkimuksen aineistolle ominaista on sen ilmaisullinen rikkaus, monitasoisuus ja kompleksisuus (Alasuutari, 2001). Alasuutarin kanssa samoilla linjoilla on Kananen. Kanasen (2014) mukaan laadullinen tutkimus tulee kysymykseen pääsääntöisesti silloin, kun ilmiöstä tiedetään etukäteen vähän. Jos ilmiöstä ei ole etukäteistietoja, teorioita, malleja tai tutkimusta, on laadullisen tutkimuksen menetelmin selvitettävä

mistä tässä on kyse. Laadullinen tutkimus antaa mahdollisuuden saada ilmiöstä syvällinen näkemys ja ymmärtää sitä (Kananen, 2014).

Tämä pro gradu on suuntautunut kvalitatiivisesti, eli tutkimussuuntaus on laadullinen tutkimus. Kananen (2019) käyttää termiä tutkimusote. Laadulliseen tutkimussuuntaukseen päädyttiin, koska tutkijalla oli tarve saada tietää mitkä ovat tekoälyn haavoittuvuuksia. Lisäksi tutkimuskysymykseen liittyvästä ilmiöstä on vain vähän tutkimusta, tietoa tai teoriaa. Tutkijan motivaatio oli saada tutkittavasta ilmiöstä syvällinen näkemys ja kuvaus.

On hyvä tunnistaa, että tutkimussuuntauksesta käytetään useita eri termejä. Kanasen (2015) käyttämä termi tutkimusote, on metodologinen kokonaisuus, jolla ongelma ratkaistaan. Otteen valinta riippuu tutkimusongelman luonteesta. Valitun tutkimusotteen mukana seuraavat tutkimusmenetelmät. Ne jakaantuvat tyypillisesti aineistonkeruu- ja analyysimenetelmiin (Kananen, 2015).

2.4 Tutkimuslaji

Hirsijärven ym. (2000 ja 2004) mukaan laadullisessa tutkimuksessa ollaan kiinnostuneita kielenpiirteistä, säännönmukaisuuksien keksimisistä, tekstin tai toiminnan merkityksen ymmärtämisestä ja reflektiosta. Nämä voivat jakautua pienempiin osiin. Esimerkiksi kielenpiirteet jakautuvat kommunikaatioon ja kulttuuriin. Nämä puolestaan jakautuvat tutkimuslajeiksi kuten sisällönanalyysi (Hirsijärvi ym., 2000 ja 2004).

Tämä pro gradu käyttää sisällönanalyysia tutkimuslajinaan. Tuomen ja Sarajärven (2009) mukaan sisällönanalyysissa pyritään luomaan tutkimusaineistosta kokonaisuus, joka painottuu teorialähtöisyyteen. Teoria liittyy tutkimuksen analyysiin ja analyysin lopputuloksiin. Jos sisällönanalyysissa halutaan painottaa analyysissa käytetyn päättelyn logiikkaa, voidaan tätä nimittää induktiiviseksi analyysiksi. Tiivistettynä voidaan todeta, että sisällönanalyysiä käyttävässä tutkimuksessa tutkittavasta ilmiöstä on jo tiedetty etukäteen jotain. Lisäksi dokumenttien hankita on ollut vapaata. Aineiston analyysi ja raportointi on tehty aineistolähtöisesti (Tuomi & Sarajärvi, 2009).

Käytettyyn tutkimuslajiin päädyttiin, koska haluttiin luoda tutkimusaineistosta esiin kokonaisuus, joka painottuu teorialähtöisyyteen. Tutkittavasta ilmiöstä oli jo tiedetty, mutta siitä oli saatavilla vain vähän tieteellistä tutkimusta ja tutkimustuloksia. Lisäksi valintaan vaikutti myös tutkijan vähäiset etukäteistiedot tutkittavasta aiheesta. Puutteita etukäteistiedoissa kompensoitiin kunnianhimolla ja ”pioneeritutkimuksen” haastavuuden rohkeana kohtaamisena.

Aineistolähtöisen tutkimuksen haastavuudesta kirjoittaa muun muassa Tuomi ja Sarajärvi (2009). Heidän mukaansa aineistolähtöinen tutkimus on haastava toteuttaa. Haastavuus tulee esille erityisesti siinä, että tutkijan on kyettävä kontrolloimaan aineistolähtöisen sisällönanalyysin tapahtumista aineiston ehdoilla. On erittäin tärkeää, että tämä ei tapahdu tutkijan omien ennakkoluulojen saattelemana. Tämä saattaa toteutua silloin, kun tutkijalla on paljon etukäteistietoa asiasta. Toinen sisällönanalyysin haasteista on se, että usein tutkija ei lopuksi

kykene tekemään tutkimuksessaan mielekkäitä johtopäätöksiä, vaan aineisto jää ikään kuin järjestetyiksi tuloksiksi (Tuomi & Sarajärvi, 2009).

2.5 Aineistonkeruumenetelmä

Tuomen ja Sarajärven (2009), Kanasen (2013) sekä Hirsijärven ym. (2004) mukaan laadullisen tutkimuksen yleisimmät aineistonkeruumenetelmät ovat kysely, haastattelu, havainnointi ja dokumentteihin perustuva tieto. Kananen (2019) täydentää tätä toteamalla, että laadullinen tutkimus voi perustua vain jo olemassa oleviin aineistoihin. Olemassa olevasta aineistosta tehdään analyysimenetelmällä tulkinta ja johtopäätös. Dokumentteja ja kerättyä aineistoa kutsutaan sekundääriaineistoksi, koska ne ovat jo olemassa (Kananen 2019).

Argumentointi on yksi keskeinen osa tutkimuksen tekemistä. Tuomen (2007) mukaan laadullisessa tutkimuksessa korostuvat aineiston keruu- ja analyysimenetelmät. Niiden esille tuominen on osa tulosten uskottavuutta. Tutkimusraportissa on tultava esille, että aineiston keruu- ja analyysimenetelmät on argumentoitu. Siitä miten tämä tuodaan esille, ollaan kirjallisuudessa montaa eri mieltä. Yksi vaihtoehto on, että tämä kirjoitetaan niin sanotusti raportin sisään. Argumentointi tulee esille raportissa siten, kuinka uskottavasti, monimuotoisesti ja pätevästi lähdeaineistoa käytetään. Argumentoinnin näkökulmasta raportissa korostuvat käytetyt lähteet, niiden merkityksellisyys aiheen kannalta ja lähdeviitteiden relevanttisuus (Tuomi, 2007).

Tässä pro gradussa aineistonkeruu on keskittynyt erilaisiin dokumentteihin, julkaisuihin, raportteihin ja julkaistuun kirjallisuuteen. Työssä ei kerätty empiiristä aineistoa, koska tuloksiin olisi voinut vaikuttaa liikaa tutkijan tai erityisesti empiirisen ympäristön ennakkoluulot. Tämä päätös perustuu aiheen ympäriltä tehtyjen aikaisempien tutkimusten ja erityisesti tutkimustulosten niukkuuteen. Aikaisemmat tieteelliset tutkimukset olisivat auttaneet tutkijaa tekemään luotettavampaa aineistoanalyysia empiirisen aineiston kanssa. Tämä ei ollut mahdollista edellä mainittujen syiden vuoksi. Tätä kokonaisuutta tarkasteltiin tutkimusraportin alussa kohdassa: Katsaus aikaisempiin tutkimuksiin.

2.6 Aineiston analyysimenetelmä

Tuomen ja Sarajärven (2009) mukaan sisällönanalyysissa etsitään tekstin merkityksiä. Sisällönanalyysilla tuotetuilla tutkimuksilla pyritään näkymättömän ymmärtämiseen. Laadullisissa analyyseissä puhutaan induktiivisista ja deduktiivisista analyyseista (Tuomi & Sarajärvi, 2009). Kanasen (2013) mukaan induktio tarkoittaa etenemistä yksittäisestä yleiseen. Tapausten avulla pyritään yleistyksiin. Laadullisen tutkimuksen induktiivisessa päättelyssä kerätään havaintoja. Näistä havainnoista tehdään yleistyksiä tai kehitetään teorioita. Etenemissuunta lähtee aineistosta, jonka takia käytetään myös nimitystä aineistolähtöinen tutkimus (Kananen, 2013).

Tuomen ja Sarajärven (2009) mukaan useimmat eri nimillä kulkevat laadullisen tutkimuksen analyysimenetelmät perustuvat tavalla tai toisella sisällönanalyysiin. Sisällönanalyysissa voidaan analysoida dokumentteja systemaattisesti ja objektiivisesti. Dokumentiksi voidaan ymmärtää kirjat, artikkelit, raportit ja mikä tahansa kirjalliseen muotoon tehty aineisto. Sisällönanalyysi sopii strukturoimattoman aineiston analyysiin. Tällöin analyysimenetelmällä saadaan tutkittavasta aineistosta esille ilmiöitä tiivistetyssä ja yleisessä muodossa (Tuomi & Sarajärvi, 2009).

Tuomen ja Sarajärven (2009) mukaan laadullisen tutkimuksen aineiston perusanalyysimenetelmä on sisällönanalyysi. Sisällönanalyysia voidaan myös käyttää tutkimuslajina, kuten tässä pro gradu tutkimuksessa on tehty. Tuomen (2007) mukaan aineistolähtöisessä analyysissa kerätystä tutkimusaineistosta pyritään luomaan teoreettinen kokonaisuus. Tämän lisäksi aineistosta valitaan tutkimuksen tarkoituksen mukaisia analyysiyksikköjä. Keskeinen ajatus on, että analyysiyksiköt eivät ole etukäteen valittuja, vaan ne nousevat keskeisiksi tutkimustyön edetessä (Tuomi, 2007). Tuomen (2007) sekä Hirsijärven ym. (2004) mukaan analyysiyksiköt voivat olla esimerkiksi sanoja tai lauseita. Tämän pro gradun analyysiyksikkönä käytettiin sanoja tekoäly, haavoittuvuus ja kybertoimintaympäristö. Lisäksi käytettiin lausetta: Tekoälyn haavoittuvuudet kybertoimintaympäristössä. Sanat ja lause käännettiin myös englanniksi.

Tässä pro gradu tutkimuksessa sisällön analyysin logiikkaa tarkasteltiin yksittäisistä tiedoista yleiseen tietoon, eli induktiivisesti. Tutkimuslajina käytettiin sisällönanalyysia, joka on samalla tutkimuksen aineiston keruu- ja analyysimetodi. Tarkennettuna analyysimenetelmä oli aineistolähtöinen sisällönanalyysi. Tutkimuksen aineiston analyysiprosessi jaettiin kolmivaiheiseksi, johon kuului aineiston pelkistäminen, ryhmittely ja käsitteiden luominen.

Tuomi ja Sarajärvi (2009 s. 108-111) kirjoittavat, että Miles ja Hubermanin (1994) mukaan aineistolähtöisen laadullisen eli induktiivisen aineiston analyysiprosessi jaetaan kolmivaiheiseksi: 1) aineiston redusointi eli pelkistäminen, 2) aineiston klusterointi eli ryhmittely ja 3) abstrahointi eli teoreettisten käsitteiden luominen. Aineiston pelkistämisessä analysoitava tieto voi olla esimerkiksi dokumentti, josta karsitaan tutkimukselle epäolennainen tieto pois. Pelkistäminen on tiedon tiivistämistä ja pilkkomista tutkimuksen kannalta oleellisiin osiin. Se perustuu tutkimustehtävään. Pelkistäminen tapahtuu etsimällä aineistosta tutkimuskysymyksiin vastaavia ilmaisuja. Seuraavassa vaiheessa klusteroinnilla luodaan pohja tutkimuksen perusrakenteelle ja poimitaan kuvauksia tutkittavasta asiasta. Viimeisessä vaiheessa, eli aineiston abstrahoinnissa erotetaan tutkimuksen kannalta olennainen tieto. Tämän valikoidun tiedon perusteella muodostetaan teoreettisia käsitteitä. Abstrahoinissa edetään kerätyn tiedon perusteella kohti teoreettisia käsitteitä ja lopulta kohti johtopäätöksiä. Tämä toteutetaan esimerkiksi yhdistelemällä dokumenteista saatuja tietoja, kunnes ne sisältävät keskeisen aineiston (Tuomi & Sarajärvi, 2009).

Tuomen ja Sarajärven (2009) mukaan näiden kolmen vaiheen kautta aineistolähtöisessä sisällönanalyysissa saadaan yhdisteltyä käsitteet, jotka vastaavat tutkimuskysymyksiin tai -tehtävään. Aineistolähtöinen sisällönanalyysi perustuu tulkintaan ja päättelyyn. Siinä edetään kohti käsitteellisempää näkemystä tutkittavasta asiasta (Tuomi & Sarajärvi, 2009).

2.7 Tutkimuksen luotettavuus

Tuomen (2007) ja Kanasen (2013) mukaan laadullisen tutkimuksen luotettavuuden arvioinnista ei ole olemassa yksiselitteistä ohjetta. Kananen (2015) lisää vielä, että laadullisen tutkimuksen luotettavuusarvioinnin käsitteistökin on kirjava. Tuomen (2007) mukaan laadullista tutkimusta, kuten muitakin tutkimuksia, arvioidaan kokonaisuutena, jolloin sen johdonmukaisuus painottuu. Myös tutkimuksen kohteen sekä tarkoituksen, tulee olla toisiinsa nähden oikeassa suhteessa (Tuomi, 2007).

Tutkimuksen luotettavuuden tarkasteluun liittyy keskeisiä käsitteitä. Kanasen (2019) mukaan tieteellisessä tutkimuksessa luotettavuutta tarkastellaan kahden käsitteen avulla. Ne ovat reliabiliteetti ja validiteetti, jotka mittaavat tutkimuksen luotettavuutta ja laatua (Kananen, 2019). Silvermanin (1993) sekä Syrjälän, Ahosen, Syrjäläisen ja Saaren (1996) mukaan laadullisen tiedon luotettavuudessa on eritoten kysymys tulkintojen validiteetista. Aineiston kohdalla validiteetti merkitsee aitoutta ja relevanttisuutta. Aineiston on oltava aitoa. Lisäksi aineiston on oltava relevanttia ongelmanasettelun taustana olevien teoreettisten käsitteiden suhteen (Silverman, 1993 ja Syrjälä, Ahonen, Syrjäläinen & Saari, 1996). Kananen (2019) puhuu reliabiliteetista tarkoittaen tulosten pysyvyyttä ja validiteetista tarkoittaen oikeiden asioiden tutkimista.

Muutkin tutkimuskirjallisuuden ammattilaiset nostavat esille reliabiliteetin ja validiteetin. Koskisen, Alasuutarin ja Peltosen (2005) mukaan tutkimuksen aineiston luetettavuutta arvioitaessa puhutaan reliabiliteetista ja validiteetista tai tutkimuksen arvioitavuudesta. Laadullisessa tutkimuksessa näihin termeihin turvaudutaan tavallisesti silloin, kun arvioidaan, voidaanko kyseiseen tutkimukseen luottaa. Kyse voi olla myös tutkimuksessa esitetyn väitteen luotettavuudesta. Mainitut käsitteet on syytä tuntea, sillä ne ovat keskeisiä tutkimuksen laadun parantamiseen tähtääviä välineitä. Käsitteet ja niihin tiivistyvä ajattelu ohjaavat tutkimuksen arviointia. Käsitteistön ja ajattelun sisäistäminen tutkimuksen alkuvaiheessa, syventävät tutkimuksen laatua ja luotettavuutta (Koskinen, Alasuutari & Peltonen, 2005).

Pysyvyyden ja oikeiden asioiden tulkitseminen painottuu laadullisen aineiston merkitysten luotettavuuteen. Syrjälän ym. (1996) mukaan laadullisen aineiston ja siitä tulkinnan avulla löydettyjen merkitysten luotettavuus riippuu kahdesta asiasta. Ensimmäiseksi miten ne vastaavat henkilöiden ilmaisuissaan (suullinen tai kirjallinen) tarkoittamia merkityksiä. Toiseksi miten ne vastaavat teoreettisia lähtökohtia (kirjallinen) (Syrjälä ym., 1996). Tämän pro gradun aineisto koostuu ainoastaan kirjallisesta dokumentaatiosta, joten suullisen ilmaisun tulkinnallisuus jäi pois.

Luotettavuus voidaan jakaa pienempiin osiin eli kriteereihin. Kanasen (2013) mukaan laadullisen tutkimuksen luotettavuuskriteerit ovat vahvistettavuus, arvioitavuus/ dokumentaatio, tulkinnan ristiriidattomuus, luotettavuus ja saturaatio. Vahvistettavuutta voidaan parantaa luettamalla aineistoa. Tulkintaa voidaan kohentaa tiedonantajalla eli informantilla. Dokumentointia voidaan parantaa perustelemalla ratkaisuja ja valintoja. Tulkinnan ristiriidattomuutta voidaan edistää käyttämällä useista eri lähteistä kerättyä aineistoa ja käyttämällä

näiden synteesiä. Lisäksi toisen tutkijan päätyessä samaan lopputulokseen, on tulkinnan ristiriidattomuus kunnossa. Luotettavuuskysymykset tulee huomioida jo työn suunnitteluvaiheessa erityisesti ennen aineiston keruun ja analyysin tekemistä. Saturaatiossa otetaan ulkoisia havaintoja tutkittavaksi niin kauan, kunnes saavutetaan kyllääntymispiste eli saturaatio (Kananen, 2013).

Tämän pro gradu tutkimuksen luotettavuuden kulmakivet olivat lähteiden synteesi, aineisto, tutkimuksen kesto, aineiston analysointi sekä tutkimusraportin luettavuus. Painoarvo oli tutkimusaineistolla ja synteesillä. Aineistoa kerätiin laajasti. Lähdeaineistossa painotettiin synteesiä monelta eri taholta. Erityisesti siinä laajuudessa, kun se oli mahdollista. Tutkimuksen kesto pidettiin tiiviinä ja sen aloitus liitettiin osaksi tutkimusseminaaria. Aineiston keräämiseen panostettiin ja tutkimuksen aineiston analyysiprosessi jaettiin kolmivaiheiseksi. Tutkimusraportin luettavuus varmistettiin oikolukijoilla.

3 TEKOÄLYN KEHITYS

Tekoälyn kehitys on vähintäänkin mielenkiintoinen. Tekoälyn aikajanalla on aloitettu tästä: Alan Turing (1950) ”Ehdotan harkittavaksi kysymystä, voivatko laitteet ajatella.” (Siukonen & Neittaanmäki, 2019 s. 139). Ja jatkettu eteenpäin: ”Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.” Tiedemiehet Stephen Hawking, Stuart Russell, Max Tegmark ja Frank Wilczek (independent.co.uk, 2014).

3.1 Tekoäly aikaisemmin

Laitilan (2019) mukaan tekoäly alkoi nousta esille 1940-luvulta alkaen. Ensimmäisenä Alan Turing ja Alonzo Church keksivät symbolisuuteen perustuvan konnektiivisen muodon, joka jäljittelee aivoja ja hermoverkkoja. Greimanin (2020) mukaan 1950-luvulla Turingin julkaisemassa Computing machinery and intelligence -artikkelissa kuvattu Turingin testi on ollut koneiden älykkyyden mittaamisen alkuteos.

Chiversin (2019) mukaan tekoälyn ensiaskeleet otettiin toisen maailmansodan jälkimainingeissa. Sodan jälkeen oli suurta kiinnostusta siihen, mitä koneet voisivat tehdä (Chivers, 2019). Laitilan (2019), Chiversin (2019) sekä Haikosen (2017) mukaan tekoälyn yhdeksi keskeiseksi henkilöksi voidaan nimetä yhdysvaltalainen tietojenkäsittelytieteen professori John McCarthy, joka esitteli termin tekoäly Darthmouth Collegen kesäseminaarissa New Hampshiressa 1956. Kesäseminaarin nimi oli artificial intelligence (AI), tekoäly. Vuonna 1956 tämä pieni tiedemiesten ryhmä: McCarthy, Minsky, Shannon ja Rochester olivat kokoontuneet miettimään sitä, miten koneita voisi alkaa opettamaan. Heidän innoittajanaan oli ollut sodan aikana Alan Turingin teoriat ja käytännön toteutukset (Laitila, 2019; Chivers, 2019 ja Haikonen, 2017).

Vahvasen (2018) mukaan 1950-luvulla ihmisen älykkyyden imitoimiseen tähtäävä tekoälytutkimus alkoi suurten odotusten saattelemana. 1960-luvulla tekoälytutkija I. J. Good oli ensimmäisiä, jotka pohtivat ihmisälyn ylittäviä koneita. Nämä ihmisälyn ylittävät koneet voisivat suunnitella itseään älykkäämpiä koneita omatoimisesti (Vahvanen, 2018).

Siukosen ja Neittaanmäen (2019) mukaan Darthmouthin seminaarista alkaen tekoälyyn liittyvät tutkimukset, sisällöt, laitteistot, ohjelmistot, ohjelmointikielet, merkitykset sekä tavoitteet ovat muuttaneet edistymisen myötä toisenlaisiksi kuin mitä ne alussa olivat. Ihmisen uteliaisuus on vienyt tekoälytukimusta eteenpäin. Laitilan (2019), Siukosen ja Neittaanmäen (2019) mukaan tekoälyn kehitykseen on kuulunut buumien syntymisiä ja sammumisia. Eri vuosikymmeninä on syntynyt aaltoja. Ne ovat pitäneet sisällään suuria edistymisiä ja takaiskuja. Symbolisesta paradigmasta on tullut ajattelun tapa. Tämä liittyy ohjelmointiin, ihmisen tuntemaan päättelyyn ja tietokoneen ytimeen. Lopulta se liittyy kaikkeen, mitä voi kehittää (Laitila, 2019 ja Siukonen & Neittaanmäki, 2019).

3.2 Tekoäly nyt

Ollilan (2019) mukaan tekoälyn kehitys on 1960-luvulta tähän päivään asti ollut sinikäyrämäistä. Tekoälyohjelman loppuraportin (2019) mukaan tekoäly on nyt saavuttanut tason, jossa se on tiedostettu maailmanlaajuisesti. Raportin mukaan useammat maat ovat sisällyttäneet tekoälyn kansallisen kilpailukykystrategiansa yhdeksi avaintekijäksi ja laatineet kansallisen tekoälystrategian. Tekoälytekniikoiden ja strategioiden muodostamisen osasyinä ovat tietokoneiden laskentatehon kasvaminen, kvanttitietokoneiden kehittäminen ja erilaisten teknologioiden yhdistyminen. Tekoäly on juuri nyt yhteiskunnalliseen teknologiakeskusteluun sopiva teema, sekä meneillään olevan teknologisen murroksen ja innovaatioiden looginen jatkumo (Ollila, 2019 ja Tekoälyohjelman loppuraportti, 2019).

Brundagen ym. (2018) ja Patelin ym. (2019) mukaan viimeaikaiset innovaatiot tekoälyjärjestelmissä ovat mahdollistaneet merkittäviä parannuksia tietokoneavusteisien tehtävien tekoon. Lisäksi valikoimaan on tullut hyödyllisiä sovelluksia. Tekoäly on siis jo nyt keskeinen osa laajasti käytössä olevia tekniikoita. Näitä tekniikoita ovat muun muassa kuvien ja videoiden tunnistus-, merkintä- ja tekstitysjärjestelmät, kuvasynteesi, sisällöntuotanto, taiteelliset työkalut sekä kuvien ja videoiden muu käsittely. Myös puhe tekstiksi ja puhe puheeksi muunnokset, kielen käännös, kielellinen analyysi, tekstisynteesi, chatbotit sekä luonnolliset kielen ymmärtämisen tehtävät käyttävät tekoälyä. Lisää lupaavia käyttökohteita ovat tekoälyjen käyttö rahoitusmallinnuksessa, automatisoidussa kaupankäynnissä, pelien pelaamisessa, itsekulkevissa ajoneuvoissa, robottiohjausjärjestelmissä, markkinointianalyyseissä, suositusjärjestelmissä, henkilökohtaisissa apulaisissa sekä hoitajien ja lääkäreiden digitaalisissa avustajissa. Kyberturvallisuuden ja digitaalisen ympäristön tehtävissä tekoälyjärjestelmille on annettu vastuu verkkovirheiden havaitsemisesta ja läpäisytestaustyökalujen käytöstä. Hakukoneissa vastuu on annettu sisällön luokitteluun, suodatukseen ja roskapostin tunnistamiseen (Brundage ym., 2018 ja Patel ym., 2019).

Edellä kuvatun mukaan tekoälyn käsillä olevat innovaatiot ovat mahdollistaneet merkittäviä parannuksia monissa digitaalisissa- ja tietokoneavusteisissa tehtävissä. Patel ym. (2019) mukaan näiden takia tekoälyjärjestelmät tuovat meille myös uusia haasteita ja riskejä. Ne mahdollistavat pahantahtoisille toimijoille uusia keinoja väärinkäyttää tekoälyn haavoittuvuuksia (Patel ym., 2019).

Patel ym. (2019) mukaan tekoälytekniikka voi aiheuttaa muutoksia ja luoda uusia riskejä nykyiselle palveluyhteiskunnalle, vaikka sitä ei tarkoituksella väärinkäytettäisi. Joillakin aloilla tekoälykkyydestä on tullut jo niin voimakas, että koulutetut mallit ovat yleisöltä piilossa mahdollisen haitallisen käytön vuoksi. Tilanne on samansuuntainen tekoälyn haavoittuvuuksien paljastamisen kanssa. Tutkijoiden on usein tehtävä kompromissi haavoittuvuuden julkistamisen kanssa. Paljastaminen avaa aina vektorin myös mahdollisille väärinkäytöksille, jos haavoittuvuutta ei ehditä tai osata korjata (Patel ym., 2019). Shevlanin ja Dafoen (2020) mielestä julkistaminen ei ole ainoa keino tekoälyn haitallisten sovellusten torjumiseksi, koska julkaisun tuloksia voidaan käyttää väärin. Parempi tapa olisi esimerkiksi sijoittaa tutkimusyhteisön jäseniä erilaisin tavoittein sisälle eri tutkimushankkeisiin. Tällöin tutkijat voisivat lisätä haitallisen käytön

ymmärtämistä muun muassa laatimalla tekoälyn käyttöä ohjaavia normeja osana tutkimushankkeita (Shevlane ja Dafoe, 2020).

Haikosen (2017) mielestä digitaalisella ohjelmoitavalla tietokoneella on jo nyt kasassa paljon yleiselle tekoälylle tarvittavia perusominaisuuksia. Tämä on keskeinen osa tekoälyn hypoteesia. Tämä tarkoittaa sitä, että nykyisen kapean tekoälyn rinnalle voidaan tulevaisuudessa luoda yleinen tekoäly sopivalla tietokoneohjelmalla. Tämä on yleisen tekoälyn tulemisen perusta. Mikäli se osoittautuu epätodeksi, putoaa digitaaliseen tietokoneeseen perustuvalta kapean tekoälyn seuraavalta askeleelta pohja (Haikonen, 2017).

3.3 Tekoälyn seuraavat askeleet

Tekoälyn seuraavista askelista puhutaan runsaasti. Erilaiset visiot sinkoilevat lähdekirjallisuudessa ja scifi-julkaisuissa. Tarkoman (2017) mukaan seuraavan kymmenen vuoden aikana tulemme näkemään tekoälyn uuden tulemisen. Tänä aikana tekoälyn tukemat digitaaliset ratkaisut kykenevät kokonaan tai osin automatisoimaan tehtäviä, joihin ennen on tarvittu ihminen. Viimeaikojen tieteelliset tulokset, suuri datamäärä, laskentatehon kasvu ja kyvykkyys laajamittaiseen hajautettuun laskemiseen johtavat tulevaisuudessa merkittävään läpimurtoon tekoälysovelluksissa. Tekoälyratkaisut kehittyvät, mutta teknologia on vielä hyvin kaukana ihmisen kognitiosta (Tarkoma, 2017).

Osa tekoälytutkijoista pohtivat tulevaisuuden määritelmiä ja käsitteitä. Henin ym. (2020) mukaan tulevaisuuden uusi käsite on ihmiskeskeinen tekoäly. Siinä korostetaan, että tekoälyn seuraava raja ei ole vain tekninen. Seuraava raja on myös humanistinen ja eettinen. Tulevaisuuden tekoälyn on heijastettava teknisesti ihmisen älykkyyden tunnusomaista syvyyttä, parannettava inhimillisiä kykyjä ja keskittyä sen vaikutusta ihmisiin (He ym., 2020).

Ollilan (2019) mukaan tekoälyn tulevaisuus on pitkälti hämärän peitossa. Tekoälyn seurauksia on vielä hankala arvioida. Tekoälyn sovelluksilla on haluttuja seurauksia, mutta myös ”ei haluttuja” -seurauksia. Mitä pitäisi ajatella näistä tekoälyn ”ei halutuista” -seurauksista? Esimerkiksi algoritmit ovat hyviä kyvyissään palvella haluttuja tarkoituksia, mutta ne voivat olla huomaamatta vääristyneitä (Ollila, 2019). Tarkoman (2017) mukaan tekoälyteknologia mahdollistaa tulevaisuudessa pitkälle menevän toimintojen analyysin, ennustamisen ja automatisoinnin. Tekoälyllä voidaan ennakoida, ennustaa vaikutukset, kääntää puhuttua/ kirjoitettua kieltä reaaliajassa, tunnistaa verkkohyökkäykset ja estää haitallinen informaatiovaikuttaminen. Jatkossa tekoälyliikenne pitää sisällään autonomisesti liikkuvat autot, rekat, junat, laivat ja lentokoneet. Nämä autonomiset kulkuvälineet muuttavat peruuntumattomasti liikennettä ja logistiikkaa. Tekoälyn ohjaamat tehtaat tuottavat tuotteita, autonominen liikenne vie tuotteet satamiin, josta robottilaivat vievät tuotteet toiselle puolelle maapalloa (Tarkoma, 2017).

Tekoälyn tulevaisuuteen liittyy positiivisia ja negatiivisia ennustuksia. Patelin ym. (2019) mukaan tekoäly on jatkossa todennäköisesti yhtä tehokas sekä hyökkäävissä, että puolustavissa tarkoituksissa. Tämä tarkoittaa sitä, että tekoäly ei tule poikkeamaan muista järjestelmistä. Patel ym. (2019) ja Floridin ym. (2018)

mukaan tekoälykontekstissa on tulevaisuudessa käynnissä ”ase - vasta-ase -kilpailu”, joka syntyy kamppailevien voimien välillä. Townsend (2018) allekirjoittaa tämän myös ja toteaa, että peruskonfliktit hyökkääjien ja puolustajien välillä eivät tule muuttumaan tekoälykontekstissa. Tekoälyjärjestelmien osalta molemmat osapuolet pyrkivät pysymään toistensa edellä ja molemmat osapuolet onnistuvat aina hetkeksi (Townsend, 2018). OP-ryhmän liiketoimintajohtaja Harri Nummelan mukaan: ”Edessä on tulevaisuus, jossa yritysten ja muiden organisaatioiden tekoälyjärjestelmät taistelevat keskenään, ja myös rikollisjärjestöt ottavat tekoälyä käyttöön.” (Siukonen & Neittaanmäki, 2019 s. 208).

Tekoälyn tulevaisuuteen liittyy myös ihmisten luottamus tekoälyä kohtaan. Halusen (2018a) mukaan luottamus perustuu usein yhteiseen historiaan ja olemassa olevan teknologian kirjaamisesta ja jakamisesta. Euroopan komission (2019) raportin mukaan luotettavuus on ennakkoedellytys sille, että ihmiset ja yhteiskunnat käyttävät tekoälyjärjestelmiä. Jos tekoälyjärjestelmät eivät osoittaudu luotettaviksi, saattaa siitä aiheutua ei toivottuja seurauksia. Nämä ei toivotut seuraukset haittaavat tekoälyjärjestelmien käyttöönottoa. Tämä epäluottamuslause voi estää tekoälyjärjestelmien sosiaalisten ja taloudellisten hyötyjen toteutumisen (European Comission, 2019). Lehdon (2019) mukaan tekoälyn laajentuessa ihmisten päivittäiseen arkeen ja yhteiskunnan elintärkeiden toimintojen aloille, tekoälyn turvallisuudesta tulee keskeinen tekijä. Ihmisten on kyettävä luottamaan tekoälyyn. Tämän luottamuksen saavuttamiseksi tekoälyn kehittäminen ja käyttöprosessit tulevat olla kunnossa. Myös tekoäly-ympäristön on oltava turvallinen. Esimerkiksi tekoälyteknologian turvallinen järjestelmäkehitysympäristö on keskeistä. Ihmisten ja organisaatioiden apuna olevien tekoälykkäiden ja autonomisuutta edistävien järjestelmien teknologian tulee olla hallittua, turvallista ja luotettavaa ensiaskelistaan alkaen. Tämä on vaade, että käyttäjät hyväksyvät ne osaksi jokapäiväistä elämäänsä (Lehto, 2019). Kuten mitä tahansa uutta kehittyvää teknologiaa, on tekoälyäkin tarkasteltava kokonaisvaltaisemmin ja kattavammin ennen sen käyttöönottoa lisäävät Hurley ja Potter (2020).

Halunen (2018a) jatkaa, että riittävää luottamusta ja kokonaisvaltaista käyttöönottoa ei tekoälyllä vielä ole. Esimerkiksi älyautojen törmätessä haluamme selvittää mitä tapahtui. Jos autot vahingoittuvat onnettomuudessa analysointikelvottomiksi, ainoat tietolähteet ovat autojen valmistajat. Luovuttaisivatko autojen valmistajat kaiken tarvittavan tiedon auton tekoälystä? Tietoja tarvitaan siihen, että tekoäly joutuu tulevaisuudessa saavuttamaan ihmisten luottamuksen. Tekoälyillä ei ole vielä luottamuksen historiaa ja juurtuneita yhteisesti hyväksyttyjä teknologioita. Miten tekoälylle voidaan rakentaa teknisiä luottamusmenetelmiä niihin tilanteisiin, joissa useat erilaiset tekoälyt ovat vuorovaikutuksessa keskenään (Halunen, 2018a)?

Tulevaisuuden tekoälytaivaalle muodostaa uhkia tekoälyn haavoittuvuudet ja singulariteetti. Brundage ym. (2018) mukaan tekoälyn haavoittuvuudet tulevat pian näkymään tulevaisuuden turvallisuusympäristössä. Tämä johtuu siitä, että tekoälyjärjestelmät yleistyvät runsaasti. Tekoälyn haitalliselle käytölle on paljon kohteita (Brundage ym., 2018). Näitä muodostuneita uhkia tummentaa puolestaan seuraava tekoälytutkija. Haikonen (2017) huomauttaa teoksessaan, että julkisessa keskustelussa on jo ennustettu tekoälyn singulariteetti.

Lähitulevaisuudessa tekoäly on ihmistä älykkäämpi ja sen jälkeen mikään ei ole entisensä. (Haikonen, 2017).

4 TEKOÄLYÄ JA KYBERTURVALLISUUTTA

Tekoäly tai mikä tahansa teknologia vaatii toimivat lähtökohdat, etukäteen pohditut käyttötarkoitukset sekä ymmärryksen mihin se kykenee ja mihin ei. Honkelan (2017) mukaan tekoälyn edellytyksiä ovat tietokone ja tietojärjestelmäkokonaisuus, jossa tekoälyä kehitetään tai käytetään. Tämän jälkeen tarvitaan idea, miten tekoälyä halutaan käyttää ja kehittää. Eli ajatus siitä mihin suuntaan asioita halutaan tekoälyn tarjoamien mahdollisuuksien avulla viedä. Tekoäly ei ole mikään irrallinen teknologia, vaan kehityksen väline. Mitä paremmin ymmärrämme mistä tekoälyssä on kyse ja mitä sillä voi tehdä, sitä paremmin voimme kontrolloida mitä tulevaisuudessa tapahtuu (Honkela, 2017). Tulevaisuuden kontrollointi on tapahduttava tekoälyn vahvuuksien ja heikkouksien, muun muassa haavoittuvuuksien ymmärtämisellä.

Tekoälyn kehittämisessä, käyttöönotoissa ja tätä kautta ymmärryksen rakentamisessa, on tärkeää jäsentää asiat realistisiksi faktoiksi. Muun muassa Jääskeläisen (2019) mukaan tekoälyn huima kehitys on synnyttänyt todellisuudestakin poikkeavia pelkoja. Erityisenä pelkona on, että on syntymässä koneiden uusi sukupolvi. Sukupolvi, joka kykenee syrjäyttämään ihmisen ja ajattelemaan itse. Näiden pelkojen takia tekoälyn käyttö ei tule olemaan ongelmatonta. Yksi suurimmista ongelmista on tekoälyn käyttöönotto (Jääskeläinen, 2019). Käyttöönottoihin liittyy keskeisesti se, mitkä instanssit ottavat tekoälyn osaksi toimintaansa. Keskeistä on ovatko yksityishenkilöt, yritykset, viranomaiset vai julkiset toimijat tekoälyjärjestelmien käyttöönoton pioneereja?

Tekoälyn kehittäminen toteutetaan samojen prosessivaiheiden mukaan, kuin minkä tahansa muu teknologia. Näiden vaiheiden yhteydessä on ryhdytty huomioimaan yhä useammin kyberturvallisuus. Tämä on oikea suunta digiajan turvallisuutta tarkasteltaessa. Lehdon (2019) mukaan tekoälyn kyberturvallisuudella on keskeinen asema tekoälyn kehittämisessä. Mikäli tekoälyn turvallisuusratkaisuja ei toteuteta parhaalla mahdollisella tavalla, siitä voi tulla vaarallinen käyttäjilleen. Tekoäly voi myös joutua kolmannen osapuolen kontrolliin. Tekoälyn kehittämisprosessissa tulee huomioida uhkien ja haavoittuvuuksien tunnistaminen. Kehittämisessä korostuu turvallinen ohjelmisto- ja teknologiasuunnittelu koko järjestelmän elinkaaren ajan. Näitä uhkia ja haavoittuvuuksia voidaan hallita oikein toteutetuilla turvallisuusratkaisuilla. Tämä on ainut oikea suunta (Lehto, 2019).

Tekoälyn kehittämisessä ja käyttöönotoissa on menty lukumäärä edellä. Niin sanotusti kaikille loppukäyttäjille on tarjolla jotakin. Stephensonin (2018) ja Patel ym. (2019) mukaan tekoälytyökaluista ja -resursseista on tullut viime vuosina helposti saatavia. Esimerkiksi julkiset pilvipalvelut tarjoavat suuria määriä resursseja loppukäyttäjille tai tekoälykehittelijöille. Dataa ja tekoälypalveluita on saatavana entistä enemmän (Stephenson, 2018 ja Patel ym., 2019). Brundage ym. (2018) lisäävät edelliseen, että kehitys ei ole pysähtymässä. Uudet tekoälymahdollisuudet lisääntyvät ennennäkemättömällä nopeudella. Tekoälytekniikoilla on jo tehty monia laajasti hyödyllisiä sovelluksia. Toiminta jatkuu kiihtyvällä tahdilla. Lukemattomia tekoälysovelluksia on kehitetty ja kehitetään parhaillaan (Brundage ym., 2019). Halusen (2019) mukaan nämä kiihtyvällä tahdilla

kehittyvät tekoälysovellukset tulevat lähtemättömäksi osaksi arkeamme. Tässä tulevassa tekoälyarjessamme kannattaa muistaa, että tekoälyjärjestelmät eivät aina toimi oikein ja niitä on mahdollista huijata. Tekoälyjärjestelmät tekevät virheitä. Ne toimivat tavoilla, jotka eivät ole aina tarkoituksenmukaisia (Halunen, 2019). Tekoälyllä toimivat järjestelmät kehittyvät. Meidän on ymmärrettävä, kuinka niitä voidaan käyttää haitallisesti. Brundagen ym. (2018) ja Bradleyn (2019) mukaan olemme innoissamme näistä kehityksen tuomista asioista. Samalla meidän tulisi olla kiinnostuneita tavoista, joilla tekoälyä voidaan käyttää väärin. Esimerkiksi kykymme tuottaa vääristettyä sisältöä on tällä hetkellä tehokkaampaa, kuin kykymme havaita onko sisältö todellista vai väärennettyä. Kokonaisuuden osalta voidaan todeta, että tekoälyn haitallinen käyttö on saanut vähemmän huomiota kuin sen hyödyntäminen (Brundage ym., 2018 ja Bradley, 2019).

Patel ym. (2019) mukaan tekoälyjärjestelmiä on käytössä monilla aloilla. Keskeisimpänä rahoitus, kauppa, tiede, armeija, terveydenhuolto, lainvalvonta, tutkimus ja koulutus. Greimanin (2020) ja Patel ym. (2019) mukaan tulevaisuudessa näillä aloilla tehdään yhä enemmän ja entistä tärkeämpiä päätöksiä tekoälyjärjestelmien avulla. Jotkut näistä päätöksistä voivat jopa johtaa muutoksiin toiminnassa ja määräyksissä. Tutkijat ovat nostaneet suureksi huolenaiheeksi kysymyksen siitä, kenen pitäisi olla vastuussa tekoälyn päätöksenteosta? On tärkeää ymmärtää, miten ja mihin perustuen tekoälyjärjestelmät tekevät päätöksiä. Toisaalta on ymmärrettävä, miten ja miksi tekoälyjärjestelmät tekevät virheitä tehdessään päätöksiä (Greiman, 2020 ja Patel ym., 2019). Asiaa tarkentaa Tarkoma (2017). Hänen mukaansa on keskeistä sisäistää, että tekoälyn autonomisten toimintoja vastaan voidaan hyökätä. Tämä aiheuttaa väärän tiedon levittämistä ja vaaratilanteita. Kun ymmärtää minkälaisia hyökkäyksiä voidaan tehdä tekoälyjärjestelmiä vastaan, voidaan huonoja ratkaisuja vähentää. Tämä minimoi tekoälyn haavoittuvuuksien käyttöä haitallisiin tarkoituksiin (Patel ym., 2019 ja Tarkoma, 2017).

Tekoälyjärjestelmät eivät poikkea perinteisestä kyberturvallisuuden kissa ja hiiri -kilpailusta. Patel ym. (2019) mukaan tekoälyn tarjoamat mahdollisuudet ovat yhtä tehokkaita sekä hyökkäävissä, että puolustavissa tarkoituksissa. Tekoälyn köydenvetokilpailu on väistämätöntä näiden kilpailevien voimien välillä. Kun tekoälyjärjestelmät yleistyvät, on luonnollista olettaa, että niitä vastaan opitaan hyökkäämään. Tarkoman (2017) ja Patel ym. (2019) mukaan jotkut tekoälyjärjestelmät ovat olleet onnistuneiden hyökkäysten kohteena jo vuosia.

Patel ym. (2019) mukaan pahantahtoisten toimijoiden puuhat tekoälyä vastaan eivät rajoitu yksittäisiin kohteisiin. Yrityksiin, viranomaisiin ja julkishallintoon kohdistuvilla toimilla on laajemmat yhteiskunnalliset vaikutukset. Toisaalta yksityishenkilöiden henkilökohtaisilla tiedoilla on tunnustettu intimiteetti ja sosiaalinen arvo. Nykyisin yksityisyyden merkitys on tunnustettu yhä enemmän välineeksi, jolla suojellaan omien arvojen kannalta merkityksellisiä tietoja (Patel ym., 2019). Euroopan komission (2019) raportin mukaan on mahdollista, että tekoälyjärjestelmät voivat tehdä ihmisen käyttäytymisen digitaalisista tallenteista päätelmiä. Päätelmät voivat koskea yksilöiden mieltymyksiä, kuten esimerkiksi sukupuolista suuntautumista ja uskonnollista tai poliittista näkemystä (European Comission, 2019). Tämä edellä mainittu intimiteettiin liittyvä

tekoälyhaavoittuvuuksien etsintä ja niihin vaikuttaminen, voi tapahtua niin yksityisellä kuin julkisellakin sektorilla huonon kyberturvallisuuden saattelemana.

Kilpailu, ymmärryksen puute ja välinpitämättömyys lisäävät uhkia myös tekoälymaailmassa. Ollilan (2019) mukaan tekoälyn käytön väärät toimintatavat synnyttävät riskejä. Inhimillisten taitojen arvo vähenee, ihmisen vastuu poistuu, yksilön kontrolli heikkenee ja ihmisen itsemääräämisoikeus heikkenee. Riskien syntyminen voi johtua väärin kohdistetuista kannustimista, ahneudesta, häijyistä aikomuksista tai geopoliittisesta kilpailusta (Ollila, 2019). Tekoälyn ollessa hypekäyrän huipulla yksi keskeinen syy on varmasti Ollilan mainitsema kilpailutilanne. Samaa mieltä on Patel ym. (2019). Heidän mukaansa yritykset tai organisaatiot luopuvat turvallisuuteen liittyvistä periaatteistaan, pyrkiessään pysymään kilpailukykyisenä. Kilpailutilanteen takia yritykset ja organisaatiot jättävät huomioimatta tietoturvallisuuden kohtia. Tämä on suuntaus kohti heikkolaatuisia ja nopeasti markkinoille tulevia tekoälyjärjestelmiä. Esimerkkinä on esineiden internet, jota pidetään ongelmallisena useimpien tietoturvallisuuden ammattilaisten keskuudessa. Samanlainen piittaamattomuus voisi olla tulevaisuuden kannalta haitallista tekoälyjärjestelmissä (Patel ym., 2019).

Järvisen (2018) mukaan tietoturva pyrkii tietojen, tiedostojen ja yksittäisten koneiden suojaamiseen. Tietoturvallisesti toimiessasi suojaat oman, perheen ja työnantajan toimintaa (Järvinen, 2018). Järvisen (2018) ja Tarkoman (2017) mukaan kyberturvallisuus tarkoittaa tietoturvan ulottamista yhteiskunnan peruspalveluihin kuten sähkön, ruokahuollon, veden jakelun, liikenteen, työ- ja talousjärjestelmän, tiedon välittämiseen sekä tietoliikenteen toimimiseen valtiollisessa mittakaavassa. Nämä edellä mainitut yhteiskunnan elintärkeät toiminnot pitävät turvallisen arjen pyörimässä (Järvinen, 2018 ja Tarkoma, 2017). Järvisen mukaan (2018) kaikki nyky-yhteiskunnan peruspalvelut toimivat tietotekniikalla ja niiden ohjaamina. Tämä tarkoittaa sitä, että pienilläkin häiriöillä voi olla kriittisiä vaikutuksia. Tietotekniikan laajamittainen käyttö on tehnyt valtioista haavoittuvaisia, joten yhteiskunnan elintärkeiden peruspalveluiden turvaaminen on tärkeää (Järvinen, 2018). Tekoäly ja kyberturvallisuus, sekä näihin turvautuvat ratkaisut liittyvät olennaisesti kokonaisturvallisuuden käsitteeseen. Ne liittyvät niihin kyvykkyyksiin, joita tarvitaan yhteiskunnan perustoimintojen ylläpitämiseen ja suojaamiseen. Lisäksi on ymmärrettävä, että tekoäly ja kyberturvallisuus eivät kytkeydy vain yhteiskunnan tai valtion sisältä tuleviin uhkiin. Järvisen (2018) mukaan verkossa ei ole maantieteellisiä rajoja ja jokainen valtio haluaa päättää omista asioistaan valtiollisten rajojensa sisäpuolella. Mikään valtio ei halua, että sen valtiollisia asioita vakoillaan tietoverkkojen kautta. Myös kansan mielipiteeseen ja -kuviin voidaan yrittää vaikuttaa tietoverkkojen kautta (Järvinen, 2018).

Tekoäly on mukana kyberturvallisuudessa niin käytännössä, kuin teknologiakeskustelussa. Kilpatrickin (2019) mukaan kyberturvallisuudesta on tullut yksi keskeinen asia teknologian alueella. Sen yksi keskeinen osa-alue on pysyä haittaohjelmien ja tietohyökkäysmenetelmien jatkuvassa kehityksessä mukana (Kilpatrick, 2019). Kilpatrickin (2019) ja Lehdon (2019) mukaan tekoälyllä on liityntä kyberturvallisuuteen. Pahantahtoiset toimijat käyttävät taitojaan haitallisesti ja ovat löytäneet tekoälystä tähän uuden työkalun. Kun kyberturvallisuus

siirtyy tekoälyn alueelle, on tärkeää olla tietoinen tekoälyyn liittyvistä uhkista (Kilpatrick, 2019 ja Lehto 2019).

Tekoälyn haavoittuvuuksien väärinkäyttöä pitää tarkastella usealta suunnalta, myös kyberturvallisuuden kannalta. Esimerkiksi Kilpatrickin (2019) mukaan kyberturvallisuusyritykset ovat tutkineet tekoälyä hyökkäysten ehkäisyn välineenä. Hakkerit ovat ajatelleet päinvastaista. Tekoäly ja sen kyky oppia voivat antaa pahantahtoisille toimijoille mahdollisuuden kiertää tyypillisiä puolustusmekanismeja. Hakkerit hakevat tekoälystä keinotekoista älykkyyttä (Kilpatrick, 2019). Hyvien tai pahojen pääsyä tekoälyapajille ei voida estää. Halusen (2019) mukaan tekoälyn ympärille tulee syntymään hakkeriyhteisö, samalla tavalla kuin kybertoimintaympäristön ympärille on syntynyt. Tekoälyratkaisuja hyödyntävien tahojen tulisi ottaa oppia aikaisemmasta kehityksestä ja toivottaa tekoälyjen heikkouksien etsijät tervetulleiksi. Näin saadaan tekoälytekniikoista parempia ja turvallisempia (Halunen, 2019).

Kuten edellä totesimme, ovat kyberturvallisuusyritykset ottaneet tekoälyn käyttöönsä. Tämä tarkoittaa sitä, että sama tietotaito ja ymmärrys on käytössä myös negatiivisessa toiminnassa. Patel ym. (2019) mukaan kyberturvallisuuden työkaluissa käytetään jo tekoälytekniikoita. Nämä työkalut ovat ilkeiden toimijoiden käytettävissä, samoin kuin turvallisuustutkijoiden ja kyberasiantuntijoiden (Patel ym., 2019). Myös Mitchellin (2018) ja Kilpatrickin (2019) mukaan kehittyneimmät kyberpuolustusteknologiat luottavat muun muassa tekoälyn oppimiseen. Tästä esimerkkinä on verkkohyökkäyksiä vastaan toimiminen. Tavallisissa virustentorjuntatyökaluissa käytetään näiden oppimisteknologioiden kautta saatuja uhkatietokantoja. Ilkeämieliset toimijat voivat hyödyntää tekoälyn oppimisprosessia. Oppimisprosessia voi sabotoida pilaamalla sen käyttämän algoritmin tai tietokannan, josta nämä järjestelmät oppivat tunnistamaan haittakoodit. Lisäämällä väärän koodin prosessiin, tekoälyjärjestelmä saadaan tuottamaan epäaitoja vastauksia. Tämä heikentää suunniteltuja toimintoja ja vähentää luottamusta tekoälyjärjestelmää kohtaan (Mitchell, 2018 ja Kilpatrick, 2019). Comiterin (2019) mielestä edellä kuvatut tekoälyä vastaan kohdistuvat hyökkäykset eroavat kyberturvallisuusongelmista. Tekoälyä vastaan tapahtuvat hyökkäykset eivät ole vain virheitä koodissa, sillä ne eivät aina vaadi edes tietokonetta (Comiter, 2019).

Kuten kyberturvallisuuteen liittyviä haavoittuvuuksia etsittäessä, myös tekoälyn haavoittuvuuksien tutkimuksessa tarvitaan avointa keskustelua sekä testausta kirjoittaa Halunen (2019). Hän jatkaa, että esimerkiksi tekoälymenetelmien arviointiin tarkoitettuja alustaratkaisuja on kehitetty vain muutamia. Tilanne poikkeaa täysin kyberkontekstista. Tekoälyn kokonaisvaltaiseen harhauttamiseen tarkoitettua avointa alustaa ei ole. Kyberturvallisuuden testaukseen tämän tyyppisiä alustoja on saatavilla internetistä avoimesti. Avoin alusta harjoitteluun ja tutkimiseen olisi syytä rakentaa pian. Tällöin olisi mahdollisuus löytää ajoissa esimerkiksi helposti harhautettavat tekoälyratkaisut. Löytymisen jälkeen voidaan tehdä tilalle turvallisempia ja parempia ratkaisuja (Halunen, 2019).

5 TEKOÄLYÄ VASTAAN KOHDISTUVAT HYÖKKÄYKSET

Useat tekoälytutkijat ovat todenneet, että tekoäly ei ole turvassa hyökkäyksiltä. Näissä offensiiveissa on samoja periaatteita ja tapoja kuin muissakin kyberympäristössä tapahtuvissa hyökkäyksissä. Comiterin (2019) mukaan tekoälyä vastaan tapahtuvat hyökkäykset ovat erityisen vaarallisia, koska hyökkäysmallit eivät tarvitse olla havaittavissa. Ne voivat olla jopa täysin huomaamattomia. Hyökkäykset voivat olla kirurgisia, muuttaen vain pienen osan tiedosta vääräksi. Lisäksi hyökkäyskuviot voivat olla ihmisen silmälle havaitsemattomia. Tämä johtuu siitä, että täysin digitaalisessa maailmassa muutokset voivat tapahtua esimerkiksi yksittäisellä pikselitasolla. Tämä tekee muutoksista niin pieniä, että ne ovat kirjaimellisesti näkymättömiä (Comiter, 2019).

Comiterin (2019) mukaan tekoälyä vastaan hyökkääminen voidaan tehdä monin eri tavoin:

- Vaurioittamisessa hyökkääjä haluaa aiheuttaa vahinkoja toteuttamalla tekoälyjärjestelmään toimintahäiriön.
- Piilottamisessa hyökkääjä haluaa estää tekoälyjärjestelmää havaitsemasta jotakin.
- Järjestelmän luottamuksen heikentämisessä hyökkääjä haluaa, että käyttäjä menettää luottamuksen tekoälyjärjestelmäänsä (Comiter, 2019).

5.1 Hyökkäysten jako

Tekoälyä vastaan tapahtuvien hyökkäysten ja muun negatiivisen toiminnan kiinnostuksen kohteita on lukematon määrä. Comiterin (2019) mukaan hyökkäyksillä pahantahtoiset toimijat voivat manipuloida tekoälyjärjestelmät toimimaan tai palvelemaan kohti heidän määrittämiään haitallisia päämääriä. Toisaalta tavoite voi olla vain aiheuttaa tekoälylle toimintahäiriö. Nämä tekoälyhyökkäykset eroavat perinteisistä kyberrikoksista. Tekoälyn taustalla on algoritmeja, entiteettejä, oppimista, kouluttamista ja fyysisiä kohteita. Nämä hyökkäykset voivat tapahtua eri muodoissa, mutta kohdistuvat erilaisiin haavoittuvuuksiin tekoälyn taustalla olevissa kokonaisuuksissa (Comiter, 2019).

Hyökkäyksiä voidaan jakaa usealla eri tavalla. Yksi keskeinen karkea jako tulee esille muun muassa Patelin ym. (2019) ja Fralickin (2019) käyttämässä jaottelussa. Patelin ym. (2019) mukaan hyökkäykset tekoälyä vastaan voidaan tehdä joko ”valkoisen laatikon” tai ”mustan laatikon” kautta. Näiden lisäksi puhutaan myös ”harmaan laatikon” kautta tehtävistä hyökkäyksistä, jotka sijoittuvat ”valkoisen laatikon” ja ”mustan laatikon” välimaastoon. Hyökkäysten jako perustuu vastustajan yhteydestä tai pääsystä järjestelmään (Patel ym., 2019).

Vähäkainun ym. (2020) ja Fralickin (2019) mukaan valkoisen laatikon hyökkäyksissä pahantahtoinen toimija tuntee tekoälyn ja sen ominaisuudet.

Ominaisuuksia ovat esimerkiksi tekoälyjärjestelmän rakenne ja parametrit. Vähäkainun ym. (2020), Fralickin (2019) ja Patelin ym. (2019) mukaan valkoisen laatikon hyökkäysmenetelmät edellyttävät, että hyökkääjällä on suora pääsy kohteeseensa. Tämä tarkoittaa sitä, että hyökkääjällä on pääsy tekoälyn koodiin, arkkitehtuuriin tai parametreihin. Näiden lisäksi voi olla pääsy myös materiaaliin, jota käytettiin tekoälyn kouluttamiseen (Vähäkainu, Lehto & Kariluoto, 2020; Fralick, 2019 ja Patel ym., 2019).

Fralickin (2019) mukaan "musta laatikko" -hyökkäyksissä pahantahtoiset toimijat eivät tunne kohteen tekoälyä eikä ominaisuuksia. Vähäkainu ym. (2020) ja Patel ym. (2019) täydentävät, että mustan laatikon hyökkäyksillä ei ole suoraa pääsyä kohteeseensa. Pääsy rajoittuu kyselyjen suorittamiseen tekoälyllä toimivaan palveluun. Kyselyjä tehdään pääsääntöisesti internetin kautta. Tällöin hyökkääjällä ei ole tietoa tekoälyjärjestelmän sisäisestä arkkitehtuurista tai tekoälyn kouluttamiseen käytetystä tiedosta. Mustan laatikon hyökkäykset toimivat suorittamalla iteratiivisia kyselyjä kohdetekoälyä vastaan. Tämän jälkeen tarkkaillaan sen antamia tuloksia tai toimintaa. Tietoa kerätään, jotta ymmärretään paremmin kohteena olevaa tekoälyä (Fralick, 2019; Vähäkainu, Lehto & Kariluoto, 2020 ja Patel ym., 2019).

Patel ym. (2019) mukaan on olemassa myös tekniikoita, jotka jäävät valkoisen ja mustan laatikon kautta tehtävien hyökkäysten väliin. Fralick (2019) nimeää tämän tyyppiset hyökkäykset ”harmaan laatikon” kautta tehtäviksi hyökkäyksiksi. Harmaan laatikon hyökkäyksissä hyökkääjä ei tiedä kaikkea kohteena olevasta tekoälystä. Patel ym. (2019) mukaan esimerkiksi standardeilla esiopetettu lähes kohdetta vastaava tekoäly, voidaan ladata pohjaksi internetistä. Tämän jälkeen hyökkääjä voi rakentaa ja jatkokouluttaa siitä mahdollisimman kohteen kaltaiseksi. Kopion rakentamisen jälkeen valkoisen laatikon hyökkäystekniikoita testataan kyseiselle kopiolle ennen lopullista hyökkäystä itse kohteeseen (Fralick, 2019 ja Patel ym., 2019).

Tekoäly-ympyröissä käytetään myös muuta merkitystä ”mustalle laatikolle”. Floridin ym. (2018) mukaan musta laatikko on mentaliteetti, jonka mukaan tekoälyjärjestelmän päätöksentekoprosessi nähdään olevan ihmisen ymmärtämisen ja siten valvonnan ulkopuolella. Siukosen ja Neittaanmäen (2019) mukaan tekoäly-ympäristössä käytössä oleva termi musta laatikko tarkoittaa sitä, että tietojärjestelmän sisään livahtaa vaikeasti löydettäviä virheitä monimutkaisten ohjelmistojen, ohjelmointien, perusoletusten ja tekniikoiden kautta. Tämä voi tapahtua vahingossa tai tahallisesti (Siukonen & Neittaanmäki, 2019).

Samalle käsitteelle on käytössä myös muita määritelmiä. MC.AI (2019) keräämän artikkelin mukaan tekoälyjärjestelmät sisältävät omistusoikeuskoodin, jota ei ole annettu yleisön saataville. Tätä kutsutaan käsitteellä musta laatikko. Ulkopuolinen tietää ainoastaan, että kyse on vain maagisesta mustasta laatikosta, joka antaa tuloksen. Miten tai miksi se sai tuon tuloksen, on liikesalaisuus (MC.AI, 2019). MC.AI:sta (2019) löytyvän artikkelin kanssa samaa kieltä puhuu Comiter (2019). Comiterin (2019) mukaan tiedämme mitä mustaan laatikkoon menee ja tiedämme mitä tulee ulos. Ongelma on se, että emme tiedä mitä näiden kahden välillä tapahtuu. MC.AI (2019) keräämän artikkelin mukaan mustan laatikon tietojen mahdollinen vääristäminen ja sen sisältämät vinoumat, voivat helposti eskaloida ongelmia. Tämä on seurausta tekoälyjärjestelmän mustan laatikon

avoimuuden puutteesta (MC.AI, 2019). Comiterin (2019) mukaan esimerkiksi mustan laatikon liikesalaisuuden alla toimivat algoritmit vaikeuttavat toiminnan tarkastamista. Tämä tekee tekoälyn toiminnan arvioinnista vaikeaa. Mustien laatikoiden takia on mahdotonta sanoa, onko tekoäly vaarantunut? Tai onko sitä vastaan hyökätty tai miksi se ei suoriudu hyvin? Tämä ominaisuus erottaa tekoälyn haavoittuvuudet kyberturvallisuusongelmista, joissa on yleensä selkeät määritelmät. Sitä mitä emme ymmärrä, emme voi korjata toteaa Comiter (Comiter, 2019).

5.2 Hyökkäysluokat tekoälyä vastaan

Fralick (2019) mukaan hyökkääjät ovat aivan yhtä ahkeria kuin puolustajat. Hänen mukaansa on olemassa monia erilaisia hyökkäysmenetelmiä ja -tapoja, joita hyökkääjät voivat käyttää edukseen (Fralick, 2019). Patelin ym. (2019) mukaan tekoälyyn perustuvien järjestelmien vastaiset hyökkäykset voidaan jakaa käsitteinä neljään pääluokkaan hyökkääjän motiivin perusteella

- luottamuksellisuushyökkäykset,
- eheyshyökkäykset,
- saatavuushyökkäykset ja
- replikointihyökkäykset (Patel ym., 2019).

Patel ym. (2019) tarkentaa omaa jakoaan niin, että hyökkääjän tavoitenäkökulmasta katsottuna saatavuushyökkäykset ovat samanlaisia kuin eheyshyökkäykset. Ne vain käyttävät erilaisia tekniikoita (Patel ym., 2019).

Comiter (2019) mukaan tekoälyyn perustuvien järjestelmien vastaiset hyökkäykset voidaan jakaa käsitteinä kahteen pääluokkaan

- myrkytyshyökkäykset ja
- syötehyökkäykset (Comiter, 2019).

Kuten kohdassa keskeiset käsitteet todettiin, tekoälykontekstin määritelmät eivät ole kaikilta osin vakiintuneet. Määritelmät saattavat olla samoja, mutta niistä puhutaan eri käsitteillä. Tähän vaikuttaa tietysti myös käsitteen kääntäminen suomen kielelle. Vakiintumattomuus tulee esille muun muassa tekoälyä vastaan tehtävien hyökkäysten luokittelemisessa. Edellä mainittujen käsitteiden määritelmää tarkasteltaessa Comiterin (2019) myrkytyshyökkäys vastaa Patelin ym. (2019) pääluokkaa eheyshyökkäykset. Comiterin (2019) syötehyökkäys vastaa Patelin ym. (2019) pääluokkaa saatavuushyökkäykset.

5.2.1 Luottamuksellisuushyökkäykset

Patel ym. (2019) mukaan luottamuksellisuushyökkäykset paljastavat tekoälyn kouluttamiseen käytetyt tiedot. Luottamuksellisuushyökkäyksiä voidaan

käyttää määrittämään, käytettiinkö jotain tiettyä tietoa tekoälyn koulutuksen aikana. Luottamuksellisuushyökkäyksien kautta saatetaan saada esille arkaluonteisia asioita tekoälystä tai sen käyttämistä tiedoista (Patel ym., 2019).

5.2.2 Eheys- tai myrkytyshyökkäys

Patel ym. (2019) mukaan eheyshyökkäykset aiheuttavat tekoälyn poikkeavan käyttäytymisen koulutustietojen manipuloinnilla. Eheyshyökkäys toteutuu esimerkiksi verkossa olevan tekoälyn opettamisella väärällä datalla. Nämä eheyshyökkäykset johtavat tekoälyn antaman informaation vääristymiseen, koska pahantahtoiset toimijat ovat myrkyttäneet tekoälyn jo koulutusvaiheessa. Toinen vaihtoehto on peukaloida koulutusmateriaaleja, millä tekoälyä opetetaan aivan alkuvaiheessa sen ollessa vielä irti verkosta. Tällöin tekoälyjärjestelmän käyttämät manipuloidut koulutusmateriaalit, saavat sen tekemään vääriä asioita taas koulutuksen jatkovaiheissa (Patel ym., 2019).

Comiter (2019) nimeää eheyshyökkäykset myrkytyshyökkäyksiksi. Myös Khurana, Mittal ja Joshi (2019) sekä Shen ja Xia (2020) käyttävät tätä määritelmää. Myrkytyshyökkäyksissä hyökkääjä pyrkii vahingoittamaan pääsääntöisesti oppimassa tai opettavana olevaa tekoälyä. Myrkytyshyökkäyksissä tekoälylle syötetään manipuloitua dataa, jonka kautta se alkaa vääristämään antamaansa tietoa (Comiter, 2019; Khurana, Mittal & Joshi, 2019; Shen & Xia, 2020). Kun tekoälyä on myrkytetty riittävästi, se muuttuu luonnostaan virheelliseksi ja sitä voidaan alkaa hallitsemaan. Tällöin tekoälyn käyttöönottoprosessia on manipuloitu niin, että tuloksena on järjestelmän toimintahäiriö hyökkääjän haluamalla tavalla.

Comiterin (2019) mukaan myrkytyshyökkäykset tapahtuvat esimerkiksi tekoälyjärjestelmän oppimisprosessin aikana. Tekoälyjärjestelmän myrkyttämiseksi hyökkääjän on manipuloitava myös käytössä olevan tekoälyn jatko-oppiminen tai lisätiedon lähde. Tämä mahdollistaa sen, että tekoäly alkaa epäonnistumaan vain hyökkääjän valitsemissa asioissa. Tietoa voidaan käsitellä monin tavoin. Yksi tapa on vääristää kelvollinen tietojoukko, vaihtamalla vain kelvolliset tiedot epäaitoon tietoon (Comiter, 2019).

Comiterin (2019) mukaan toinen vaihtoehto myrkytyshyökkäyksen toteuttamiseksi, on hyökätä tietojoukkojen keräämisprosessiin. Keräämisprosessissa tietoja hankitaan koulutusmateriaaliksi. Tämä vääristää tekoälyä tehokkaasti alusta alkaen. Tämä johtaa siihen, että tekoälyn käyttäjät eivät voi enää luottaa ovatko heidän keräämänsä tiedot oikeita. Manipuloitujen tietojen löytäminen voi olla hyvin vaikea tietoaineistojen laajuuden vuoksi (Comiter, 2019).

Shenin ja Xianin (2020) sekä Comiterin (2019) mukaan kolmas tapa toteuttaa myrkytyshyökkäys, on vääristää tekoälyn käyttämää algoritmia. Hyökkäys voidaan toteuttaa esimerkiksi niin kutsutulla Troijan hevosella. Hyökkääjät voivat manipuloida käynnissä oleva algoritmia tekoälyn antaman informaation vääristämiseksi (Shen ja Xian, 2020 ja Comiter, 2019).

5.2.3 Saatavuus- tai syötehyökkäys

Patel ym. (2019) mukaan saatavuushyökkäykset aiheuttavat tekoälyyn poikkeavan käyttäytymisen, estämällä siltä olennaisten tietojen saamisen. Nämä saatavuushyökkäykset johtavat käytössä olevan tekoälyn antaman informaation vääristymiseen. Tämä on mahdollista, koska hyökkääjät ovat manipuloineet tietoa tekoälyn tarvitsemaa tietoa. Ihmiselle tieto näyttää muuttumattomalta, mutta tekoälylle väärä tieto näyttää täysin erilaiselta kuin oikea tieto. Saatavuushyökkäyksillä pahantahtoinen toimija tekee häiriöitä ympäristöön. Ympäristöä voi muokata niin, että tekoäly luokittelee väärin ympärillä olevat esineet tai niistä saatavan tiedon (Patel ym., 2019).

Comiter (2019) nimeää saatavuushyökkäykset syötehyökkäyksiksi. Niissä hyökkääjät voivat luoda muutoksia kohteeseen, joka huijaa tekoälyjärjestelmän tekemään virheen. Esimerkiksi muuttamalla kuviot ristiriitaan tietojoukon kanssa. Tässä hyökkäysmuodossa on hyvä tunnistaa, että tekoälyjärjestelmää vastaan suunnattu syöttöhyökkäys ei aina tarvitse tietokonetta. Syötehyökkäyksessä syötteen käsitteleminen muuttaa järjestelmän antamaa tulosta hyökkääjän haluamalla tavalla. Nämä syötehyökkäykset laukaisevat tekoälyjärjestelmän toimintahäiriön muuttamalla sitä, mitä järjestelmään tarjotaan. Syötehyökkäykset eivät edellytä sitä, että hyökkääjä olisi päässyt vioittamaan itse tekoälyjärjestelmää. Huipputekniset tekoälyjärjestelmät ovat erittäin tarkkoja niiden tietoteknisten syötteiden eheydestä (Comiter, 2019).

Comiterin (2019) mukaan syötehyökkäysmuodot voidaan jakaa kahdella tavalla: havaittavuus ja muoto. Havaitseminen on ominaista, jos hyökkäys on havaittavissa ihmissilmälle. Muoto kuvaa sitä, onko hyökkäysvektori fyysisen reaalimaailman esine vai digitaalisen maailman ominaisuus (Comiter, 2019).

5.2.4 Replikointihyökkäys

Patel ym. (2019) mukaan replikointihyökkäyksissä vastustaja yrittää kopioida kohteen tekoälyä tai sen käyttämiä tietoja. Replikointihyökkäyksien tavoitteena on saada riittävästi tietoa kohteesta, jotta siitä voidaan luoda kopio. Luotua kopiota voidaan käyttää omiin harjoittelutarpeisiin tai hyökkäyksiin alkuperäistä tekoälyjärjestelmää vastaan (Patel ym., 2019). Comiter (2019) nostaa esille myös tekoälyn jäljittelemisen ja takaporttien luomisen tekoälyjärjestelmiin omassa luokittelussaan.

5.2.5 Tekoälyn korvaaminen toisella versiolla

Comiterin (2019) mukaan viimeinen tapa hyökätä tekoälyä vastaan on yksinkertaisesti korvata se vääristetyllä versiolla. Tämä voidaan tehdä esimerkiksi perinteisen kyberhyökkäyksen turvin. Ensin on saatava varastettua kohdetekoäly mallipohjaksi, jos sitä ei ole kyetty tekemään itse. Tämän jälkeen malliksi tehty tekoäly koulutetaan omin toimenpitein hieman omanlaiseksi. Lopuksi se asennetaan alkuperäisen kohdetekoälyn tilalle uutena versiona. Vaikka tekoäly olisi opetettu oikein, oikealla tietoaineistolla, varustettu aukottomilla algoritmeilla ja

tarkistettu toiminnan alussa perusteellisesti, voidaan se silti korvata toisella tekoälyversiolla (Comiter, 2019). Korvattu versio tekoälystä on lähtökohdiltaan ja perusolemukseltaan kuten alkuperäinen, mutta se on vääristynyt.

6 TEKOÄLYN HAAVOITTUVUUDET

Uusista tekoälyjärjestelmistä on tulossa nykyisiä elinvoimaisempia ja laaja-alaisempia. Tähän elinvoimaisuuteen ja laajenevaan tekoälypohjaiseen tukeen liittyy myös uhkia. Hyvä esimerkki uhkista ovat tekoälyn haavoittuvuudet. Kun tekoäly on tukenamme yhä moninaisimmissa prosesseissa, sen olisi hyvä toimia kaatumatta, toimintahäiriöittä, virheettä ja hakkeroitumatta. Tällöin siihen luotetaan ja sen toimintavarmuus on hyvä. Jos näin ei ole, tekoälyjärjestelmän kaatuminen, toimintahäiriö ja hakkeroimisen mahdollisuus voi johtua sen haavoittuvuuksista.

Vakaan ja tietoturvallisen tekoälyn vaade on käsillä. Tegmarkin (2018) mukaan tekoälyjärjestelmiä on kaatunut ja niitä tulee vielä kaatumaan. Tämä ei ole hyvä asia, sillä tekoäly on jo siirtynyt tutkijoiden kammioista tosimaailmaan. Tämän päivän maailmassa epävakaa ja huonosti kovennettu tekoälyjärjestelmä voi kaataa päivittäisen liikkumisen, sähkönjakelun, teollisuuden tai osakemarkkinat (Tegmark, 2018). Samasta asiasta kirjoittaa myös Tarkoma. Tarkoman (2017) mukaan tekoälyä hyödynnetään yhä laajenevassa määrin edellä mainituilla yhteiskunnan elintärkeillä osa-alueilla. Tekoälyratkaisut ovat eri toimintojen rakenneosia ja mahdollistavat entistä laajemman automatisoinnin, sekä tuen päätöksentekoon. Samalla tekoälyjärjestelmät luovat uusia haasteita järjestelmien suojaamiseen ja toiminnan varmistamiseen. Tämä on jokapäiväistä realismia, sillä tekoäly on osa automatisoidumpaa yhteiskuntaa (Tarkoma, 2017). Tegmarkin (2018) mukaan mitä automatisoidumpi yhteiskunta ja mitä tehokkaampi hyökkäävä kyky ovat, sitä tuhoisimmaksi tähän liittyvät kyberuhat voivat muuttua. Jos pystyy hakkeroimaan itseohjautuvat autot, autopilotilla lentävät lentokoneet, ydinreaktorit, teollisuusrobotit, viestintäjärjestelmät, rahoitusjärjestelmät tai voimaverkot, voi horjuttaa haluamaansa kohdetta (Tegmark, 2018). Nämä kaikki ovat uhattuina tekoälyn haavoittuvuuksien kautta.

Monet muutkin tekoälyn kanssa toimivat ovat huomanneet tekoälyn nopean edistymisen monilla tärkeillä rintamilla. Erityisesti sen läsnäolon osana arkipäivän askareita. Tegmark (2018) kirjoittaa, että tekoälyllä on ollut jo pitkään melko tasaista edistymistä. Nyt näyttää, että tekoälyn nopea kehitys jatkuu todennäköisesti vuosia. Ei ole perusteltavaa syytä sille, miksi edistyminen ei juuri nyt jatkuisi. Edistyminen tulee olemaan lähes keskeytymätöntä, kunnes tekoäly nousee ihmisen kykyjen tasolle. Tekoäly voi nousta jopa ihmisen kyvyn ohi useissa tehtävissä. Tämä on vääjäämätöntä, sillä nykyisin kaikki omassa sivilisaatiossa rakastamamme on ihmisen älyn tuotetta. Jos pystymme vahvistamaan nykyistä kehitysastetta tekoälyllä, saisimme mahdollisuuden kehittää ja parantaa nykyistä tasoamme. Tällöin jopa vaatimaton tekoäly voisi johtaa merkittäviin parannuksiin tieteessä, teknologiassa ja turvallisuudessa (Tegmark, 2018). Halunen antaa esimerkin nykyisen kehitystasomme parannuksista. Halusen (2019) mukaan tekoälyjärjestelmät pystyvät jo nyt huomattavasti ihmistä parempaan tunnistamiseen ja ennustamiseen monilla alueilla. Muun muassa kasvojentunnistuksessa ja huulilta lukemisessa tekoälyjärjestelmät ovat edistyneitä. Jotkut tekoälysovellukset tulkitsevat jopa mikroilmeitä. Tunnistamisessa tekoälyjärjestelmät eivät tee virheitä väsyneenä tai huolimattomuuttaan (Halunen, 2019). Näihin tekoälyn tekemiin tunnistuksiin liittyy myös haasteita. Niissä mahdollisesti

esiintyviä haavoittuvuuksia voidaan käyttää myös negatiivisessa mielessä. On vastattava oikein moniin erilaisiin kysymyksiin, ennen kuin voimme nauttia tekoälyn kehityksen eduista luomatta uusia ongelmia Tegmark (2018) painottaa.

Tarkoman (2017) mukaan tekoälyjärjestelmissä on haavoittuvuuksia. Niitä on esimerkiksi sen toiminta- tai toteutustavan kautta. Haavoittuvuuksien avulla pahantahtoinen toimija voi hyökätä tekoälyä vastaan. Hyökkäys voi tapahtua kohdejärjestelmästä tehdyn mallin avulla. Mallilla voi simuloida tilanteita ja selvittää kohteen heikot kohdat (Tarkoma, 2017). Hyökkääjä voi mallintaa kohteen ja harjoitella sillä kohdetekoälyn haavoittuvuuksien hyödyntämistä tai vain vastapuolen tekoälyn harhauttamista. Tarkoman ajatuksia täydentävät Comiter (2019) ja Patel ym. (2019). Heidän mukaansa tekoälyjärjestelmän haavoittuvuuksien kautta tapahtuvat hyökkäykset voivat olla vain välillisesti vahingollisia ja täten vaikeasti havaittavissa. Tämän takia tekoälyn haavoittuvuuksien vahingollista käyttöä vastaan on vaikea puolustautua (Comiter, 2019 ja Patel ym., 2019).

Tekoälyllä on perustarpeita ja tietyt asiat on tehtävä ennen kuin sen palveluista päästään hyötymään. Tekoälyä pitää muun muassa opettaa, jotta se saadaan toimimaan loppukäyttäjän haluamalla tavalla. Opetustapahtuma on tärkeä vaihe tekoälyn opettamisprosessissa, mutta samalla se on yksi vaaranpaikka. MC.AI (2019) keräämän artikkelin mukaan tekoälyjärjestelmälle on annettava oppimista koskevat säännöt ja parametrit. Kieli ja logiikka toimivat ihmiselle, mutta eivät tekoälylle. Ne ovat riittämättömiä työkaluja, koska tekoäly tarvitsee toimiakseen enemmän. Tekoälyn on haastavaa omaksua sitä ympäröivää maailmaa, joka pitää sisällään paljon erilaisia ilmiöitä (MC.AI, 2019).

Tämä tekoälyn omaksuminen, eli oppiminen, ei toteudu itsekseen. Siihen on käytettävä koulutusmateriaalia. MC.AI (2019) keräämän artikkelin mukaan tekoäly on toisaalta vain niin hyvä kuin sen opettamiseen käytetty koulutusmateriaali. Tieto on tässäkin kohtaa valtaa. On monia tekijöitä, jotka voivat uhata koulutusmateriaalin eheyttä. Voi syntyy tilanteita, joista ei ole vertailukelpoista tietoa. Tällöin tekoäly voi joutua umpikujaan (MC.AI, 2019). Patel ym. (2019) tarkentavat, että suurin osa tekoälyjärjestelmistä on tällä hetkellä opetettu vain nyt saatavilla olevan koulutusmateriaalin mukaan.

MC.AI (2019) keräämän artikkelin mukaan oppimisen ja koulutusmateriaalin lisäksi myös tekoälyjärjestelmien algoritmit edustavat "täydellisiä malleja" maailmasta. Ei ole siis ilmeistä, miten tekoäly selviää monimutkaisessa maailmassa ja sen epäselvissä tilanteissa (MC.AI, 2019). Myös Townsend antaa oman kantansa tekoälyn algoritmeihin, koulutusmateriaaliin ja oppimiseen. Townsendin (2018) mukaan tekoälyn oppiminen toimii, mutta kahdella varauksella. Tekoälyn oppimisen laadukkuus riippuu oppimisalgoritmin laadusta ja koulutusmateriaalin eheydestä. Mahdollinen väärinkäyttö voi tapahtua molemmilla alueilla, eli algoritmin manipuloinnissa tai koulutusmateriaalin vääristämisessä. Jälkimmäisessä myrkytetään datajoukko, josta tekoäly oppii (Townsend, 2018).

Tekoälyllä on myös koodi- ja ohjelmistovirheiden, laitetason sekä tunnistustekniikoiden haavoittuvuuksia. Nämä yhdessä oppimisen, opetuksen ja koulutusmateriaalin haavoittuvuuksien kanssa, saavat tekoälyjärjestelmät vaikuttamaan monimutkaisilta. Patelin ym. (2019) mukaan suurin osa tekoälyjärjestelmistä ovat kuitenkin melko yksinkertaisia ja toimivat lukuisissa samantyylisissä skenaarioissa. Toisin on silloin, kun tekoälyjärjestelmiltä haetaan enemmän

ihmismäisen ymmärryksen kaltaista toimintaa. Patel ym. (2019) mukaan tekoälyjärjestelmät toimivat yksinkertaisissa tehtävissä suurimman osan ajasta hyvin. Kun mallit muuttuvat monimutkaisemmiksi, on tekoälyn ymmärrykseen liittyvien haavoittuvuuksien vähentäminenkin paljon vaikeampaa (Patel ym., 2019).

Tekoälyn tapa ymmärtää tai olla ymmärtämättä, on osa sen haavoittuvuutta. Tätä ymmärryksen puutetta voidaan yrittää käyttää hyödyksi huijaamalla. Muun muassa Halusen (2018b ja 2019) ja Lehdon (2019) mukaan tekoäly on altis tulemaan huijatuksi. Lehto (2019) tarkentaa, että viime vuosina tutkijat ovat havainnollistaneet monia keinoja huijata tekoälypohjaisia järjestelmiä. Tämä on ollut mahdollista esimerkiksi hyödyntämällä niiden taipumusta havaita datasta erilaisia säännönmukaisuuksia (Lehto, 2019). Halunen (2019) jatkaa, että kuvaa tai ääntä tunnistava järjestelmä voidaan harhauttaa luulemaan henkilöä tai esinettä täysin toiseksi. Tekoälyn ymmärrys ei yleensä riitä tunnistamaan tätä huijaamista (Halunen, 2019).

Edellä kuvatun mukaisille tekoälyn haavoittuvuuksille ei tule heti olemaan oikeanlaista tai oikeanaikaista laastaria. Tämä johtuu hyvin monesti taloudellisesta aspektista, mutta myös tekoälyn sovelluskehittäjistä tai loppukäyttäjistä. Taloudellinen kilpailu ei useinkaan mahdollista tunnollisesti ja hartaudella tehtyjä tekoälyratkaisuja. Laarin toim. (2019) mukaan uudessa teknologiassa laiminlyödään usein tietoturvaominaisuuksien viimeistelty kehittäminen. Tämän takia ne ovat houkuttelevia kohteita kyberhyökkäyksille (Laari toim., 2019). Näiden keskeneräisten ratkaisujen markkinoille tuleminen vesittää myös tekoälyn haavoittuvuuksien etsinnän.

Halusen (2019) mukaan tekoälyjärjestelmien haavoittuvuuksia kannattaisi etsiä ajoissa, koska käynnissä on kilpajuoksu. Kilpajuoksu näkyy monilla muillakin ”uusilla” aloilla, kuten esimerkiksi ICT- ja kyberturvallisuusaloilla. Uusia järjestelmiä luodaan nopeasti, jolloin sekä vanhoista että uusista löydetään haavoittuvuuksia. Näitä haavoittuvuuksia yritetään paikata mahdollisimman nopeasti jälkijättöisesti. Tämä ei toteudu aina niin hyvin ja joutuisasti, kuin olisi tarpeen (Halunen, 2019). Halusen (2019) kanssa samaa mieltä on Tekoälyohjelman loppuraportti ja Bradley. Tekoälyohjelman loppuraportin (2019) ja Bradleyn (2019) artikkelin mukaan tekoälyn väärinkäytön mahdollistavat haavoittuvuudet on tunnistettava ajoissa. Tämä toteutetaan esimerkiksi tekoälysovellusten kehittäjien kautta. Kehittäjien tulee huolehtia turvallisuudesta ja varautua väärinkäytöksiin jo suunnitteluvaiheesta lähtien (security by design). Kaikkien tekoälyä kehittävien organisaatioiden riskienhallinnan on oltava aktiivisesti tietoisia lähestymistapojensa puutteista. Haitallisten aikomusten estäminen on oltava sisällytetty itse kehityssuunnitelmaan. Toisaalta tekoälyjärjestelmien käyttäjien koulutukseen tulee myös panostaa. Heidän on ymmärrettävä tekoälysovellusten toiminta sekä rajoitteet (Tekoälyohjelman loppuraportti, 2019 ja Bradley, 2019). Kaikki eivät ole yhtä rakentavalla kannalla. Comiterin (2019) mukaan tekoälyn haavoittuvuudet ovat haastavia paikata. Esimerkiksi yrityksen käyttämän tekoälyn haavoittuvuutena voivat olla räätälöidyt algoritmit tai alkuperäisen tekoälyversion koulutusmateriaalin vääristymät. Näihin ei auta vahva IT-osasto ja 90 merkin salasanat (Comiter, 2019).

6.1 Tekoälyn opettamiseen ja oppimiseen liittyvä haavoittuvuudet

Ollakseen hyödyksi, tekoälyn pitää tehdä jotain tuottavaa. Ennen tätä, tekoälyn on opittava tekemään. Tämän takia sitä opetetaan. Jääskeläisen (2019) mukaan tekoälyjärjestelmien rakentamisen keskeinen vaihe on niiden opettaminen ja oppiminen. Opettamisessa on keskeistä järjestelmiin syötettävän datan ominaisuuksien tunnistaminen (Jääskeläinen, 2019). Jääskeläisen kanssa samoilla linjoilla on Järvinen (2018). Hänen mukaansa tekoäly opetetaan tai sen annetaan oppia itsekseen jokin tietty tehtävä. Tekoälystä tehdään tuottava ja hyödyllinen juuri siinä, mihin se on opetettu. Tekemän siirtoja shakissa, pelaamaan go-peliä, kissavideoiden tunnistamiseen tai kielen kääntämiseen. Tekoälyä ei opeteta tekemään näitä kaikkia. Tekoäly opetetaan tekemään aluksi vain yhtä asiaa. Tekoälyn oppiminen on siis alakohtaista. Tekoälyn virittäminen uuteen tehtävään vaatii aina uuden opettamista. Esimerkiksi jos syötteessä tapahtuu jotain odottamatonta, tekoäly ei osaa reagoida siihen. Syynä voi olla se, että sitä ei ole siihen opetettu. Lisäksi oppimista ei aina voida automatisoida (Järvinen, 2018).

Tekoälyn opettaminen lähtee alakohtaisesta tarpeesta, johon se perustaa ensiaskeleensa. Ensiaskeleet painottuvat opettajansa näkemyksiin ja saatavilla oleviin pohja- sekä ennakkotietoihin. Honkelan (2017) mukaan tekoälyjärjestelmä perustuu niihin käsityksiin tai ennakkoluuloihin, joita sen kehittäjillä on. Opetusvaiheessa järjestelmälle syötetään kaikki saatavilla olevat tapaukset. Opetettujen tapauksien pohjalta tekoälylle syntyy malli, joka perustuu käytettyihin algoritmeihin ja taustalla olevaan tilastolliseen teoriaan. Lopulta tekoäly on oppinut matkimaan sille opetettuja tapauksia. Uusia tapauksia tarkastaessaan tekoäly soveltaa oppimisprosessissaan oppimaansa ja tarjoaa niiden perusteella tuloksiaan (Honkela, 2017).

Näistä tekoälyn opettamisprosessin osa-alueista on käytössä erilaisia käsitteitä ja niiden määritelmiä. Vakiintunutta tekoälyn oppimisen taksonomiaa ei ole, mutta toisaalta määritelmät ovat lähellä toisiaan. Kirjallisuudessa käytetään koneoppimisen ympäristöön liittyviä käsitteitä, jolloin oppimisprosessin osa-alueet ovat ymmärrettävissä terminologisista vivahteistaan huolimatta. Jälleen kerran myös kielelliset ja käännöstekniset vivahde-erot tulevat esille.

Ollilan (2019) mukaan tekoälyn oppiminen voidaan jakaa kolmeen osa-alueeseen: ohjattuun, ohjaamattomaan ja vahvistusoppimiseen. Myös Honkela (2017) käyttää edellä mainittua jakoa. Patel ym. (2019) jakavat oppimistekniikoita ohjattuun, puoliohjattuun, valvomattomiin ja vahvistusoppimiseen. Heidän mukaansa tekoälyjärjestelmien opettamiseen käytetyt tekniikat riippuvat ongelmatilanteesta ja käytettävissä olevasta tiedosta (Patel ym., 2019).

Honkelan (2017) mukaan tekoälyjärjestelmä on yhteydessä ympäröivään maailmaan. Ohjatussa oppimisessa tekoälyä ohjataan kädestä pitäen. Tällöin opetukseen tarvitaan paljon tapauksia. Honkela jatkaa, että tekoälyjärjestelmä oppii tietoa ympäröivästä maailmasta ja antaa informaatiota toimintansa tuloksena (Honkela, 2017). Ollilan (2019) mukaan ohjattua oppimista voidaan käyttää, kun tiedetään mitä opetusdatasta halutaan saada tulokseksi. Ohjatussa oppimisessa järjestelmä ottaa vastaan syötettä ja antaa vastauksiksi tuloksia (Ollila,

2017). Patelin ym. (2019) mukaan ohjattuja oppimistekniikoita käytetään kouluttamaan tekoälyjärjestelmiä täysin merkityksellisistä tiedosta. Lisäksi ohjattua oppimista voi tarkentaa puoliohjattuun oppimistekniikkaan. Puoliohjattuja oppimistekniikoita käytetään kouluttamaan tekoälyjärjestelmiä, jossa tieto on osittain merkityksellistä (Patel ym., 2019).

Honkelan (2017) mukaan ohjaamattomassa oppimisessa ei anneta valmiita vastauksia. Syötteeksi annetaan materiaalia tai keinotekoisia kokemuksia, mutta ei valmista tietoa (Honkela, 2017). Ollilan (2019) mukaan ohjaamatonta oppimista voidaan käyttää, kun valmiita luokkia ei ole vielä olemassa tai kun data voidaan esittää havainnollisessa kuvallisessa muodossa. Honkela (2017) jatkaa, että ohjaamaton oppiminen ei ole ihmisen valinnoista täysin riippumatonta. Ihminen määrittää parametrit ja muuttujat, jotka vaikuttavat lopputulokseen. Muuttujina on esimerkiksi asioita, joilla käsiteltäviä asioita kuvataan (Honkela, 2017). Patel ym. (2019) mukaan valvomattomia oppimistekniikoita käytetään täysin esikäsittelemättömän tiedon kanssa.

Honkelan (2017) mukaan vahvistusoppimisessa tekoäly saa toiminnastaan jatkuvasti palautetta. Palautteen kautta tekoäly saa tietää, millä tavalla sen tekemä ratkaisu toimii. Aluksi ratkaisut voivat olla kömpelöitä. Ajan myötä se kehittyy koko ajan taitavammaksi, jos puitteet ovat sopivia (Honkela, 2017). Ollilan (2019) mukaan vahvistusoppiminen liittyy tilanteeseen, jossa keinotekoisen toimijan pitää pystyä operoimaan monimutkaisessa ympäristössä. Ympäristö antaa toiminnasta negatiivista tai positiivista palautetta. Tekoäly pyrkii ratkaisuun, joka tuottaa positiivista palautetta eniten (Ollila, 2019). Kokkarisen (2003) mukaan vahvistusoppiminen on yksi mahdollisuus opettaa tekoälyä toimimaan siten, että se saavuttaa tavoitteensa mahdollisimman hyvin. Erityisesti silloin, kun järjestelmän rakentaja ei itsekkään osaa etukäteen sanoa optimaalista toimintoa. Vahvistusoppimisessa tekoäly oppii omista kokemuksista. Oppimistehtävästä tekee hankalan se, että toimintojen tulokset ja tilanteista saatavat tiedot voivat olla satunnaisia ja viiveellisiä (Kokkarinen, 2003). Patelin ym. (2019) mukaan vahvistusoppimisen tekniikoita käytetään kouluttamaan tekoälyjärjestelmiä vuorovaikutukseen ympäristöjen kanssa.

Nämä tekoälyn oppimisprosessit eivät ole turvassa pahantahtoisilta toimijoilta. Oppimisprosessin haavoittuvuudessa on suoria liityntäpintoja esimerkiksi koulutusmateriaaliin. Comiterin (2019) mukaan tekoälyn oppiminen on yksi sen haavoittuvuus. Tekoälyn oppii esimerkiksi kuviosta, jotka on helppo hajottaa. Tekoäly oppii tilastotietoja, joita on suhteellisen helppo vääristää (Comiter, 2019). Comiterin kanssa samaa mieltä on Halunen. Halusen (2018b) mukaan erilaisia opetusvaiheen peukaloinnin tapoja on useita. Jääskeläisen (2019) ja Vahvasen (2018) mukaan tekoälyn oppimiseen liittyy haavoittuvuuksia, koska opetusvaiheessa ei välttämättä osata tunnistaa tai ottaa huomioon kaikkia tosielämän mahdollisia tilanteita. Tämä antaa mahdollisuuden harhauttaa tekoälyjärjestelmää tavalla, joka ei menisi läpi ihmiseltä. Tekoäly oppii esimerkiksi asioita, joita sen ei olisi alun perinkään haluttu oppivan. Vaikka tekoäly tarkkailisi dataa ja mukautuisi saamiinsa signaaleihin, se lopulta tekee juuri niin kuin ihminen on sen ohjelmoinut (Vahvanen, 2018 ja Jääskeläinen, 2019).

Kuten edellä todettiin, tekoälyn oppiminen on yksi sen haavoittuvuuksista. Kun haavoittuvuus on löydetty tai tunnistettu, on tarkoituksenmukaista estää

sen väärinkäyttö. Tämä ei ole tekoälyn oppimisen kohdalla yksinkertaista. Patel ym. (2019) mukaan tekoälyn oppimisen kohdistuvia hyökkäyksiä on vaikea estää. Oppimisvaiheessa on olemassa monia tapoja opettaa tekoäly tuottamaan vääriä tuloksia. Monimutkaiset ongelmat voidaan ratkaista vain käyttämällä tekoälyn opettamiseen hienostuneita oppimismalleja. Tekoälyä on kuitenkin vaikea opettaa kaikkia mahdollisia hyökkäyksiä vastaan (Patel ym., 2019).

Tekoäly pystyy omaksumaan asioita, mutta sen opettaminen ja oppiminen sisältää haavoittuvuuksia. Vahvasen (2018) mukaan tekoälyn kehittäminen ei ole enää sitä, että ihminen ohjelmoi koneeseen kaikki mahdolliset toimintaperiaatteet. Jatkossa tekoälyn on pystyttävä tähän yhä omatoimisemmin (Vahvanen, 2018). Kaikki eivät ole asiassa vielä yhtä positiivisia. MC.AI (2019) keräämän artikkelin mukaan tekoäly tarvitsee vielä paljon ihmisen valvontaa, jotta se voidaan opettaa oikein. Ilman ihmisen valvontaa tekoälyjärjestelmät tekevät usein sellaisia virheitä, joita ihmiset eivät tekisi. Omatoiminen tai valvomaton oppiminen on tekoälyjärjestelmien tavoite. Tämän tavoitteen saavuttaminen kestää vielä hetken (MC.AI, 2019).

6.2 Tekoälyn koulutusmateriaalin haavoittuvuudet

Tapamme kuvailla maailmaa on osa tekoälyn koulutukseen käytettävää materiaalia. Tämä data saattaa mahdollisesti sisältää vinoutumia, vääristymiä ja harhoja. Näiden ei haluttujen tietojen on mahdollista jalkautua osaksi tekoälyn toimintaa osana sen käyttöönottoprosessia. Prosessi, joka lähtee liikkeelle olemassa olevasta datasta ja tästä muokatusta tiedosta. MC.AI:sta (2019) löytyvän artikkelin mukaan tekoälyn koulutusmateriaali voi olla täynnä vääristymiä. Nämä yksinkertaisetkin harhat tai vinoumat ovat väistämättömiä asioita. Meillä kaikilla on niitä. Ne ovat upotettuja tapaamme kuvailla maailmaa. Useimmat niistä ovat hyvänlaatuisia ja loogisia, mutta eivät kaikki. Esimerkiksi tekoälyn koulutusmateriaalia on voitu kerätä aikaisemmilta ajanhetkiltä. Tällöin se ei välttämättä sovi nykyhetkeen (MC.AI, 2019).

Järvisen (2018) mukaan vinoutumat ja harhat vääristävät ihmisen kykyä hahmottaa maailmaa, vastaanottaa informaatiota ja käsitellä aiemmin tapahtunutta. Halunen on Järvisen kanssa samaa mieltä. Halusen (2019) mukaan tekoälyjärjestelmän suunnittelussa ja opetusvaiheessa on huomioitava opetusdatan mahdolliset vääristymät. Jos näin ei tehdä, järjestelmä alkaa käytännössä toteuttaa niitä (Halunen, 2019). Järvisen (2018) mukaan tutkijat ovat löytäneet erilaisia tapoja, joilla ihmisen aivot vääristelevät havaintoja ja niiden perusteella tehtäviä johtopäätöksiä. Vinoutumia ja havaintovääristymiä ovat muun muassa vahvistusharha, ankkurivaikutus, strutsiefekti, innovaatioharha, haloefekti, selviytymisharha, ylivertaisuusharha ja sokean pisteen harha (Järvinen, 2018). Myös Ollilan (2019) mukaan kaikenlaisen datan käyttö altistaa vääristymille. Inhimillinen erehtyväisyys, moraalinen ja kognitiivinen ajatusmaailma on mukana ihmisen ohjaamassa datan keruussa. Tätä kautta se siirtyy lopulta esimerkiksi tekoälyn tekemään analysointiin tai päätöksentekoon. Ihmiset päättävät, miten data

varastoidaan, miten toimintoja luokitellaan ja kvantifoidaan. Ihmiset myös tulkitsevat tekoälyn käyttämää dataa ja sen tekemiä toimia (Ollila, 2019).

Zhengin (2017) mukaan ensisijainen menetelmä tekoälyn vaarantamiseksi on ollut tietojen vääristäminen. Zhengin kanssa samaa mieltä on Halunen. Halusen (2018b) mukaan tekoälyn haavoittuvuuksiin liittyy koulutusmateriaalin peukalointi tai manipulointi. Vaarantuminen tapahtuu esimerkiksi jo tekoälyjärjestelmän opetusvaiheessa, joissa käytetään vääristynyttä koulutusmateriaalia. Opetusvaiheessa tekoäly pyrkii oppimaan tai sitä opetetaan haluttuun tehtävään mahdollisimman hyvin koulutusmateriaalin perusteella. Tekoäly ei toimi halutun mukaisesti, jos pahantahtoinen toimija pääsee käsiksi tähän koulutusmateriaaliin. Hän voi muokata koulutusmateriaalia itselleen edulliseksi tai tekoälyn tehtävään liittyen epätarkoituksenmukaisemmaksi. Toisaalta tekoälyn koulutusmateriaalia voi muuttaa vain vähän ja mahdollisimman huomaamattomasti. Tämä pienikin muutos mahdollistaa toteuttajan haluaman lopputuloksen ennemmin tai myöhemmin (Halunen, 2018b). Jos koulutusmateriaalin manipulointia ei havaita, mikä tahansa organisaatio voi joutua pulaan lisää Zheng (2017).

Tekoälyjärjestelmät vaativat runsaasti laaja-alaista koulutusmateriaalia opetusvaiheeseensa liittyen. Opetusvaiheen koulutusmateriaali tulee olla riittävän laadukasta minimalistisine puutteineen ja virheineen. Lehdon (2019) mukaan koulutusmateriaalin manipulointiriskit liittyvät juuri tekoälyn käyttämään runsaaseen data-aineistoon. Tarkoma (2017) laajentaa tätä toteamalla, että tekoäly voi lisäksi erehtyä puutteellisen koulutusdatan johdosta. Lisäksi se voi erehtyä uuden ennakoimattoman tilanteen takia. Toiminnan konteksti rajoittaa siis tekoälyn toimintaa. Jos tähän reagoidaan väärin, kontekstin muuttumisella voidaan aiheuttaa vain lisää virheellisiä päätelmiä. Lopulta ajaudutaan käyttämään eri yhteydessä opittua mallia (Tarkoma, 2017). Lehto on samoilla linjoilla Tarkoman kanssa. Lehdon (2019) mukaan data-aineistojen manipulointi tuottaa virheellisiä lähtötietoja tekoälylle. Virheelliset lähtötiedot eskaloituvat virheellisiksi toiminnoiksi. Esimerkiksi koulutusmateriaalin tai datan saastuttaminen onnistuu yksinkertaisesti niin, että aineistoon lisätään ”jotain” ylimääräistä. Ihminen ei tätä huomaa. Tekoäly huomaa ja tekee vääriä tulkintoja valvovan ihmisen huomaamatta. Ytimekkäästi voidaan todeta, että tekoälyjärjestelmät tekevät vääriä johtopäätöksiä vääristetystä, saastutetusta tai manipuloidusta datasta (Lehto, 2019).

Tieto on keskeistä tekoälylle. Erityisesti tekoälyn opetukseen käytetty tieto. Edellisten kappaleiden perusteella voidaankin sanoa, että tekoälyn yksi haavoittuvuus liittyy sen koulutusmateriaaliin. Lisäksi on todettava, että tekoälyn koulutusmateriaaliin liittyvät haavoittuvuudet ovat jo täällä. Muun muassa Järvisen (2018) mukaan nykyisissä tekoälyjärjestelmissä on huomattu ongelmia, jotka vääristävät niiden toimintaa. Tämä on seuraus siitä, että tekoäly on ihmisten antamien koulutusmateriaalien varassa. Esimerkiksi tekoälyn loppukäyttäjä ei voi tietää, millaisella materiaalilla tekoälyjärjestelmä on lopunperin koulutettu (Järvinen, 2018).

Tätä koulutusmateriaalihaavoittuvuutta pohtii myös Comiter. Comiterin (2019) mukaan tekoälyllä on riippuvuus tiedoista. Tekoäly oppii purkamalla malleja annetuista tietoaineistoista. Toisin kuin ihmiset, tekoälyllä ei ole perustietoja. Ihminen hyödyntää omassa oppimisessa perustietoja. Tekoäly ei tätä tee.

Tekoälyn kyky riippuu täysin sen saamista tiedoista. Tekoäly on tämän takia erittäin riippuvainen koulutusmateriaalista. Opetukseen käytetty aineisto on tekoälyn keskeisin tietolähde. Jos joku on peukaloinut tai manipuloinut informaatiota, näistä tiedoista opittu toiminta on virheellistä. Pahantahtoiset toimijat voivat korruptoida opetukseen käytettävää tietojoukkoa ja estää näin tekoälyä oppimasta tiettyjä toimintoja. He voivat jopa asentaa koulutusmateriaalin kautta salaisia takaportteja, joita voidaan käyttää tekoälyä vastaan joskus tulevaisuudessa. Kun pohjatyöt on koulutusmateriaalin kanssa tehty, voidaan haavoittuvuutta käyttää silloin kun aika on otollisin (Comiter, 2019).

Koulutusmateriaalin vääristämiseen liittyy myös muita uhkakuvia. Halusen (2018b) mukaan opetusvaiheen jälkeen tekoälyn ei pitäisi pitää enää sisällään yksittäisiä koulutusmateriaalin osasia. Esimerkiksi yksilöivien tietojen perusteita, yrityssalaisuuksia tai henkilötietoja (Halunen, 2018b). Myös Zhengin (2017) mukaan tekoälyn kouluttamiseen on käytetty valtavia määriä arkaluontoisia tietoja. Nämä tiedot voivat olla esimerkiksi henkilöiden yksityiselämästä. Băjenescun (2018) mukaan hakkerit, jotka haluavat varastaa henkilötietoja tai luottamuksellisia tietoja kohdistavat kiinnostuksensa yhä useammin tekoälyjärjestelmiin.

Halusen (2018b) mukaan tutkimuksissa on esitetty menetelmiä, joilla tekoälyn algoritmista on saatu ulos arkaluontoista koulutusmateriaalia. MC.AI keräämä artikkeli tuo esille saman asian. MC.AI:sta (2019) löytyvän artikkelin mukaan tekoälyjärjestelmän kouluttamiseen tarvittava tiedonkeruu itsessään voi osoittautua vaaralliseksi. Tämä tapahtuu erityisesti silloin, jos organisaatiossa on liikesalaisuuksia. Tekoälyjärjestelmä kerää näihin liikesalaisuuksiin liittyvää tietoa, kehittyäkseen vielä paremmaksi tekoälyjärjestelmäksi. Tämä toiminta ei ole täysin tietoturvallista. Tekoäly tai sen taustalla oleva henkilö, saattaa pystyä käyttämään saatavillaan olevia tietoja väärin. Tämä voi tapahtua tiedostaen tai tiedostamatta (MC.AI, 2019). Myös Järvisen (2018) mukaan keskitetty tietovarasto on riski. Jos joku hakkeroi tiensä rekistereihin, hän voi tehdä suurtakin vahinkoa. Tietojen tuhoaminen on nykyisin vahingoiltaan vähäisin, sillä tiedot voidaan palauttaa varmuuskopioilta. Toisaalta pitkään huomaamattomana jatkunut tietojen vääristäminen on voinut korruptoida myös varmuuskopiot (Järvinen, 2018).

Tekoälyn oppiminen vaatii jatkuvasti valtavia määriä koulutusmateriaalia ja tietoja. Tietoja tarvitaan itse asiasta ja suoritettavien tehtävien ympäriltä. Esimerkiksi MC.AI (2019) keräämän artikkelin mukaan palvelusopimusten ehtojen ja tietosuojakäytäntöjen hyväksymisistä kysytään verkossa jatkuvasti. Laitteet ja verkkosivustot keräävät sinulta tietoja aina. Omiin tarkoitusperiisi sopivien asioiden nopeampi löytyminen, tapahtuu oman yksityisyytesi kustannuksella. Loppukäyttäjiltä kerättyjä tietoja ohjataan muun muassa tekoälyn käyttöön, kehittämiseen ja seuraavien tekoälyjärjestelmien koulutusmateriaaliksi (MC.AI, 2019).

6.3 Tekoälyn algoritmien haavoittuvuudet

Sautoyn (2019) ja Fryn (2018) mukaan algoritmit ovat usean modernin keksinnön takana ja ne ohjaavat digitaalisen ajan elämäämme paljon. Fryn (2018) mukaan

nämä näkymättömät koodinpalaset muodostavat nykyisen aikakauden koneiston. Algoritmit ovat antaneet maailmalle kaiken sosiaalisen median syötteistä hakukoneisiin ja satelliittinavigoinnista musiikin suosittelujärjestelmiin. Algoritmit ovat yhtä lailla osa modernia infrastruktuuriamme kuten sillat, rakennukset ja tehtaat (Fry, 2018).

Ollilan (2019) mukaan algoritmi on yksityiskohtainen kuvaus tai ohje siitä, miten tehtävä tai prosessi on suoritettava. Algoritmi on järjestykseen laitettuja yksiselitteisiä toimintoja, jotka määrittelevät lopputulokseen johtavan prosessin (Ollila, 2019). Honkela (2017) liittää algoritmikuvauksen tekoälyyn. Hänen mukaansa tekoälyn yhtenä perusteena ovat algoritmit, jotka ovat tekoälyjärjestelmien rakennuspalikoita. Ne ottavat vastaan niille annettua dataa ja muokkaavat sitä eri tavoilla. Honkela jatkaa, että tekoälyjärjestelmä ei koostu pelkästään yhdestä vaan useammasta algoritmista (Honkela, 2017). Kilpatrick (2019) ja Zheng (2017) lisäävät, että algoritmien kautta tekoäly voi osoittaa älykästä käyttäytymistä. Lehdon (2019) mukaan algoritmeja ovat muun muassa sumea logiikka, säätöjärjestelmät, geneettiset algoritmit ja parveilualgoritmit.

Fry (2018) mukaan algoritmit voidaan jakaa kahteen eri tyyppiin. Ensimmäinen algoritmityyppi perustuu sääntöihin. Sen ohjeet ovat ihmisten luomia, suoria ja yksiselitteisiä. Sääntöihin perustuvia algoritmeja on helppo ymmärtää. Tämä on myös niiden huono puoli. Ne toimivat vain sellaisissa ongelmissa, joihin ihmiset osaavat kirjoittaa ohjeet. Toinen algoritmityyppi on koneoppiva algoritmi, joka liittyy tekoälyyn. Siinä koneeseen syötetään dataa, annetaan tavoite ja palautetta. Kun kone pääsee oikeille jäljille, sen annetaan ratkaista itse paras tapa saavuttaa tavoite. Koneoppivat algoritmit ovat hyviä käsittelemään ongelmia, joissa ohjeluettelon kirjoittaminen ei toimi. Haittapuolena on, että ihminen ei välttämättä pääse selville koneen tekemästä ratkaisuprosessista. On siis todettava, että algoritmi ei ole hyvä tai paha sellaisenaan. Kyse on siitä, miten niitä käytetään (Fry, 2018).

Fry (2018) mukaan algoritmeja käytetään neljässä eri pääluokassa. Usein näitä mainittuja algoritmien käytön pääluokkia yhdistellään käytännön toimenpiteen toteuttamiseksi:

- Priorisointi on käskyjen mukaisen luettelon laatimista.
- Luokittelu on kategorioihin jakamista.
- Yhdistäminen on sidosten löytämistä.
- Suodattaminen on tärkeiden asioiden seulomista (Fry, 2018).

Kuten aikaisemmin todettiin, tekoäly tarvitsee runsaasti laadukasta ja oikeaa dataa. Kilpatrickin (2019) ja Zhengin (2017) mukaan algoritmit tarvitsevat valtavasti tietoja toimiakseen oikein ja tarkasti. Vaikka on mahdollista tehdä ennusteita ilman runsasta dataa, ison datamäärän ja algoritmien kautta saadaan täsmällisempää tietoa ja enemmän tarkkuutta. Tekoäly oppii algoritmiensa avulla. Käytön myötä tekoäly tulee taitavaksi havaitsemaan erilaisia asioita oppimiensa tietojen perusteella. Ongelmia tulee silloin, kun joku löytää tavan muuttaa algoritmia tai tietoja (Kilpatrick, 2019 ja Zheng, 2017).

Townsend (2018) mukaan on toistuvasti osoitettu, että algoritmeilla on haavoittuvuuksia. Niihin sisältyy haavoittuvuuksia, kuten väärän luokituksen

indusoiminen tai myrkytetyt koulutusmateriaalit. Samaa mieltä Townsendin kanssa on Zheng (2017). Hänen mukaansa algoritmin tietolähteen tai koulutusmenetelmän tunnistaminen on arvokasta tietoa. MC.AI (2019) artikkelin mukaan, tekoälyllä on vaikeuksia erottaa todellinen tieto väärennöksestä. Pahantahtoiset toimijat voivat pyrkiä tarkoituksella vääristämään algoritmia ja sen jatkuvaa oppimista. Tämä vääristäminen voidaan tehdä räikeästi tai hienovaraisesti (MC.AI, 2019). Lopulta algoritmit vahvistavat jo olemassa olevaa harhaa ja tuottavat tuloksia, jotka vahvistavat aikaisempia vääristyneitä tuloksia tiivistää Băjenescu (2018). Näin oravanpyörä on valmis ja kukaan ei osaa sanoa milloin homma olen lähtenyt väärille raiteille.

Ollilan (2019) mukaan algoritmien kehittäjien puolueellisuus saattaa myös siirtyä osaksi sen toimintaperiaatteita. Tämä toiminta voi olla myös tiedostamatonta. Informaatioalgoritmit ovat alttiita käyttäjien manipuloinnille. Datalähteet saattavat olla vääristyneitä ja sisältää puolueellisuutta, jolloin koneen tulkinta vääristyy ja tehdyt päätökset ovat vääriä (Ollila, 2019).

Näiden edellä mainittujen uhkien lisäksi, algoritmit ovat avoinna myös perinteisille haavoittuvuuksille. Townsendin (2018) mukaan nämä uhat ovat jo nyt olemassa. Ilkeämieliset alkavat jatkossa käyttää omia kykyjään lisätäkseen kohteeseensa kohdistuvien hyökkäysten nopeutta ja tarkkuutta (Townsend, 2018). Tähän antaa esimerkin Zhengin (2017), jonka mukaan esimerkiksi suodattimet on opetettu luokittelemaan tietyn sisältöiset sähköpostit roskapostiksi. Haitalliset toimijat voivat kiertää tämän suodatuksen sekoittamalla algoritmin kyvyn rinnastaa sisältö roskapostiin, jolloin sähköposti voi ohittaa roskapostisuodattimen (Zheng, 2017). Tekoälyn avulla voidaan sitten räätälöidä nämä sähköpostit kohdehenkilöiden mielenkiintojen mukaan kirjoittaa Korolov (2017).

Toisen esimerkin antaa Mitchell (2018), jonka mukaan kasvojen tunnistusalgoritmeja vastaan on kehitetty suodatin. Suodatinta voidaan käyttää kuvien yksityisyyden suojaamiseksi. Suodatin muuttaa kuvan pikseleitä. Muutokset ovat ihmisen silmälle lähes havaitsemattomia, mutta eivät tekoälylle (Mitchell, 2018). Konkreettinen esimerkki tästä on Chicagon yliopiston SAND-laboratorion (2020) kehittämä Fawkes-algoritmi ja ohjelmistotyökalu. Se antaa yksilölle mahdollisuuden muuttaa esimerkiksi internettiin laittamia valokuviaan. Fawkes tekee kuviin pieniä, ihmissilmälle näkymättömiä pikselitason muutoksia. Näitä muutoksia kutsutaan kuvan peittämiseksi. Peittämisen jälkeen kuvia voi käyttää tavalliseen tapaan, vaikka sosiaalisessa mediassa. Ero on siinä, että perinteinen kasvojentunnistus ei onnistu kuvien kautta (SAND Lab, 2020). Toinen tähän liittyvä esimerkki on muun muassa Mikrobitin (2019) julkaiseman artikkeli Tencent Keen Security Labin tekemistä tutkimuksista. Tencent Keen Security Labin (2019) -raportin mukaan sen tietoturvatutkijat manipuloivat autopilotilla ohjattua autoa tiehen kiinnitettävillä tarroilla. Tarrat olivat ihmiselle lähes näkymättömiä, mutta algoritmeille ne kertoivat kaistan kaartuvan. Autopilotti päätti kaartaa vasemmalle, vaikka tosiasiassa tie jatkui suoraan (Tencent Keen Security Lab, 2019).

Algoritmien kautta tekoälylle pyritään saamaan ihmismäisiä toimintoja. Tämän tavoittelu pitää sisällään algoritmeihin kohdistuvia riskejä. Lehdon (2019) mukaan algoritmiriskit liittyvät annettujen ja opittujen algoritmien perusteisiin. Algoritmeilla pyritään mallintamaan ihmisen ajattelua ohjelmallisin keinoin. Algoritmisuunnittelun haasteena on luoda toiminteita, tilanteita ja olosuhteita.

Näitä ovat muun muassa ihmismäinen looginen ja analyyttinen päätöksenteko. Nämä pitäisi toteuttaa ilman koodausvirheitä. Suunnittelun epäonnistuminen voi jopa itsessään toimia virheellisesti tai suunnitteluperusteiden vastaisesti. Algoritmiriskien haavoittuvuudet perustuvat niiden toimintatapaan, eikä ihmismäiseen älykkyyteen (Lehto, 2019). Myös Ollilan (2019) mukaan algoritmien haasteet ovat ihmismäiset. Laadukas data voi heijastaa ihmiselle ominaista käytöstä ja lajityypillisiä asenteita. Se sisältää ihmisen paheet ja moraalittomat teot. Tällöin lopputulos on teknisessä mielessä oikea, mutta käytännössä vääristynyt (Ollila, 2019).

Algoritmit kehittyvät myös tekoälymaailmassa. Mitchellin (2018) mukaan 2010 luvulla algoritmit olivat ihmisen määrittelemiä. Nykyisin algoritmit oppivat itse. Nykyisin algoritmeille ei tarvitse toimittaa muuta kuin koulutusmateriaalia (Mitchell, 2018). Tästä jatkaa Siukonen ja Neittaanmäki (2019) joiden mukaan algoritmeista ei käy enää helposti ilmi, miksi tai missä kohtaa tekoäly on tehnyt huonon päätöksen. Usein tekoälypäätösten perusteita ei anneta julkisuuteen tai niitä ei pystytä edes selvittämään (Siukonen & Neittaanmäki, 2019).

6.4 Tekoälyn koodi- ja ohjelmistovirheiden haavoittuvuudet

Tekoälyn on ennustettu olevan usean eri alan uusi visio. Tekoälyä räätälöimällä ja jatkokehittämällä voidaan saavuttaa erilaisissa ympäristöissä uusia kehityspolkuja. Siukosen ja Neittaanmäen (2019) mukaan tekoälyllä on satoja eri sovelluksia. Sovellusten kaupallinen menestys on tuonut lisävirtaa tekoälyn hyödyntämiseen ja asioiden jatkokehittämiseen. Tekoäly voi toimia esimerkiksi tukena kyberuhkien havaitsemisessa, ratkaisemisessa ja torjumisessa. Useimmat tekoälyratkaisut automatisoivat toimenpiteitä ja helpottavat ihmisen toimintaa (Siukonen & Neittaanmäki, 2019).

Sadat erilaiset tekoälysovellukset leviävät laajalle. Tällöin taustalla olevat ohjelmistot ja ohjelmointikoodit leviävät. Vääriin käsiin levinneet tiedot johtavat lopulta tekoälyjärjestelmien ohjelmistojen ja ohjelmointikoodien haavoittuvuuksien löytymiseen. Muun muassa Halusen (2018b) mukaan tekoäly voidaan yksinkertaisimmillaan ymmärtää tavallisena ohjelmistona. Ohjelmistossa on aina haavoittuvuuksia, heikkouksia ja bugeja. Haavoittuvuuksia ja heikkouksia hyödyntämällä voidaan toteuttaa tietomurtoja ja kyberhyökkäyksiä (Halunen, 2018b).

Kun mennään kohti automatisoidumpaa yhteiskuntaa, sen kybertoimintaympäristössä vaikuttavien yksittäisten haavoittuvuuksien ja heikkouksien määrä kasvaa teknologioiden lisääntyessä. Laarin toim. (2019) mukaan pelkästään laitteistojen määrä ja monimutkaisuus tuottaa loppukäyttäjille haasteita laitteiden päivitysten ajantasaisina pitämisessä. Tämä tarkoittaa sitä, että päivittämättömiä teknologioita ja järjestelmien haavoittuvuuksia pyritään hyödyntämään rikollisessakin toiminnassa (Laari toim., 2019). Tämä sama uhka on tekoälyjärjestelmien toiminnassa sen koodien ja ohjelmistojen kautta. Lehdon (2019) mukaan nämä koodi- ja ohjelmistotason riskit konkretisoituvat. Ohjelmistot ovat ihmisen tekemiä ja siksi haavoittuvia (Lehto 2019). Pagen, Bainin ja Mukhlishin

(2018) mukaan ohjelmistovikoja on vaikeampi tunnistaa ja korjata, jos ohjelma on kehitetty käyttäen jonkinasteista keinotekoista oppimista.

Lehdon (2019) ja Patelin ym. (2019) mukaan tekoälyjärjestelmä voi olla alttiina myös koodivirheiden kautta avautuville haavoittuvuuksille. Pahantahtoiset toimijat hyödyntävät bugeja ja pääsevät niiden kautta järjestelmään. Tämä tapahtuu kuten missä tahansa tietokoneohjelmassa (Lehto, 2019 ja Patel ym., 2019). Lehto (2019) jatkaa, että tekoälyjärjestelmien koodimanipulointi on todennäköistä. Tämä on potentiaalinen uhka, koska uusien sovellusten ohjelmakoodia julkaistaan usein avoimena lähdekoodina internetissä. Julkaisun jälkeen sitä vasta sovelletaan johonkin varsinaiseen käyttötarkoitukseen. Avoin lähdekoodi antaa mahdollisuuden tutkia ohjelmakoodin muokkaamista. Jos muokattua koodia saadaan tekoälyjärjestelmään, toimii sovellus vastoin alkuperäistä tarkoitusta. Tämä tarkoittaa sitä, että järjestelmät ja laitteet eivät toimi halutulla tavalla (Lehto, 2019).

Tekoälyllä on haavoittuvuuksia sen koodi- ja ohjelmistovirheiden kautta. Tämä johtuu Zhengin (2017) mielestä esimerkiksi siitä, että haavoittuvuuksia esiintyy luonnostaan koodeissa. Ratkaisuja tämän haavoittuvuuden pienentämiseksi ja poistamiseksi on jo tekeillä. Hänen mukaansa yrityksillä on tiukat menettelyt sovellusten testaamiseksi ja korjaamiseksi. Tyypillisesti yritykset julkaisevat ohjelmistonsa prototyypin beta-testaajien ryhmälle virheiden löytämiseksi ja korjaamiseksi ennen julkaisua. Toinen tapa on käynnistää palkkio-ohjelmia. Palkkio-ohjelmissa tutkijoita ja hakkereita houkutellaan ilmoittamaan virheistä tarjoamalla taloudellista palkkiota löydettyjä haavoittuvuuksia vastaan (Zheng, 2017).

6.5 Tekoälyn laite- ja komponenttitason haavoittuvuudet

Tekoälyllä on haavoittuvuuksia myös laitetasolla. Laitetason haavoittuvuudet voivat johtaa esimerkiksi tekoälyä käyttävien järjestelmien hallinnan menettämiseen. Pagen, Bainin ja Mukhlishin (2018) mukaan toimintahäiriö on riski, josta useimmat tekoälyjärjestelmien suunnittelijat ovat hyvin tietoisia. Toimintahäiriö voi johtua esimerkiksi laitteistovioista (Page, Bain & Mukhlish, 2018). Havaittu häiriö voi olla tahaton tai tahallinen. Lehdon (2019) mukaan tekoälylaitteiden laitetason hallinnan tai halutun toimintatavan muuttuminen toteutetaan laite- ja komponenttitasoilla. Laitetason haavoittuvuudet perustuvat tekoälyjärjestelmien ohjelmistoriskeihin. Jos tekoälyn käyttämiä laitteistoja ei kyetä hallitsemaan, syntyy uhkia. Tällöin järjestelmät eivät ole enää täysin käyttäjänsä kontrollissa. Pahantahtoinen toimija voi hyödyntää laitetason haavoittuvuuksia järjestelmässä, jos esimerkiksi tekoälyjärjestelmien kybersuojautuminen on toteutettu heikosti. Jos joku saa tekoälylaitteet kontrolliinsa, tämä hallinnan menetys aiheuttaa uhkia usealle muulle tasolle. Laite- ja komponenttitasolle tehdyt takaportit mahdollistavat laitteiston hallinnan mikroprosessoritasolla. Tämä voi esiintyä esimerkiksi kytkimien etähallinnoimisena. Laitteistotroijalaiset asennetaan siruihin, jolloin toiminto voidaan kytkeä päälle ennalta asetettuun tai ulkoisesti käskettyyn toimintamoodiin. Tämä toteutetaan esimerkiksi mikrosirujen

mekaanisella käsittelyllä tai ylimääräisillä komponenteilla, jotka muuttavat tekoälylaitteiston alkuperäisen idean mukaista toiminnallisuutta (Lehto, 2019).

6.6 Tekoälyn tunnistustekniikoiden haavoittuvuudet

Halusen (2018b) mukaan yksi tekoälyn haavoittuvuus on sen käyttämät tunnistustekniikat. Tekoäly pyrkii jonkinlaiseen tunnistamiseen ja tähän perustuvaan päätöksentekoon. Pahantahtoinen tekijä voi tietyissä tilanteissa harhauttaa tekoälyä. Esimerkiksi hahmontunnistuksen ja kasvojentunnistuksen alueella tähän haavoittuvuuteen on törmätty. Tekoälyalgoritmeja on harhautettu luokittelemaan asioita täysin toiseksi mitä ne ovat. Tämä on mahdollista, koska tekoälyn tunnistus perustuu hyvin erilaisiin menetelmiin. Tekoälyn tekemät tunnistusvirheet näyttävät ihmisistä käsittämättömiltä, eivätkä kaikki tunnistuksen harhautus- tai hämäysmenetelmät toimisi ihmisiä vastaan. Automatisoidussa ympäristössä, jossa tekoäly tekee päätökset ihmisen puolesta hämäykset voivat mennä läpi (Halunen, 2018b).

Comiterin (2019) mukaan on erityistä, onko hämäys havaittavissa ihmissilmälle. Hämäys voi olla fyysisen reaalimaailman esine tai digitaalisen maailman ominaisuus. Fyysisen reaalimaailman esine voi olla esimerkiksi peiteteippi tai värimuunnos. Digitaalisen maailman ominaisuus voi olla esimerkiksi kuvan päällä oleva pikselitason suodatinkuva tai häiritsevä kohina. Nämä vääristymät voidaan lisätä kuvaan suoraan tai vasta jälkikäteen (Comiter, 2019).

Väärien tunnistustulosten tuottamiseen Comiter antaa hyvän esimerkin. Comiterin (2019) mukaan tekoälykontekstissa tarvitaan yksinkertaisimmillaan vain teippiä ja muutama pieni nauhanpala. Näillä tarvikkeilla itse ajavan auton mielestä stop-risteys on muuttunut tasa-arvoiseksi risteykseksi. Tekoälylle tehtävät muutokset voivat siis olla suuria, mutta ihmiselle täysin näkyvissä. Tällöin ne on tehty näyttämään siltä, että ne sopivat täydellisesti ympäristöön tekoälyn mielestä (Comiter, 2019). Toisaalta hyökkäykset voivat olla myös täysin näkymättömiä ihmissilmälle tarkentavat Patel ym. (2019) ja Chivers (2019). Sautoy (2019) tarkentaa vielä, että hyökkäykset voivat olla ”näkymättömiä” myös tekoälylle.

Seuraava esimerkki liittyy tekoälyn käyttämään tunnistustekniikkaan, jolla esimerkiksi sosiaalisen median palvelut pystyvät tunnistamaan ihmisiä suoraan valokuvista. Tätä samaa tekniikkaa voi myös väärinkäyttää. Tivin (2018) mukaan torontolaisen yliopiston tutkijat päättivät kehittää teknologiaa, jolla kasvojentunnistusta voisi estää tunnistamasta kasvoja valokuvista. Projekti toteutettiin kilpailuttamalla kahta tekoälyä vastakkain. Toinen tekoäly yritti tunnistaa kasvoja kuvista. Toinen tekoäly pyrki tekemään kuviin pienenpieniä muutoksia, jotka estäisivät tunnistamisen. Testissä käytetyistä 600 kasvokuvasta tunnistettiin aluksi 100 prosenttia. Toisen tekoälyn tekemän kuvakäsittelyn jälkeen tunnistusprosentti oli vain 0,5 prosenttia. Muutokset olivat ihmissilmälle käytännössä näkymättömiä, mutta pystyivät harhauttamaan tekoälyä (Tivi, 2018).

6.7 Tekoälyn ymmärryksen haavoittuvuudet

MC.AI (2019) keräämän artikkelin mukaan tekoäly ei ymmärrä syy-yhteyttä. Ihmiset osaavat tunnistaa ongelmat ja miettiä mistä ne ovat lähtöisin. Jopa nuori lapsi kysyy jatkuvasti: Miksi? Tekoäly on melko hyvä vastaavuussuhteiden löytämisessä, mutta ei niinkään syy-yhteyden löytämisessä. Tekoäly löytää kuvioita vastaavia muotoja, mutta se ei ymmärrä yhteyttä löytämistään kuvioista. Jos käyttäjä ei tunnista tätä rajausta, voivat tulokset olla mitä vain (MC.AI, 2019). Sautoy (2019) mielestä tämä tarkoittaa sitä, että tekoäly ei osaa reflektoida omaa tuotostaan. Onko antamani tuotos hyvä vai huono (Sautoy, 2019)? Beedham (2020) antaa tähän liittyvän esimerkin the nextweb-verkkolehdessä julkaistussa artikkelissaan. Artikkelin mukaan israelilainen yliopisto selvitti, että autonomista autoa voi hämätä yksinkertaisilla kikoilla. Hämäyksessä lennokki heijasti omalla projektorillaan auton eteen kaksiulotteisen kuvajaisen tiellä olevasta ihmisestä, autosta, moottoripyörästä tai tien sivussa olevasta liikennemerkistä. Heijastetut kuvat saivat autopilotilla kulkevat autot reagoimaan. Tilanteessa ihminen olisi huomannut pelkän optisen illuusio, mutta autonominen auto ei tätä päättelyä osannut tehdä (Beedham, 2020). Tekoälykuskin päätöksentekoon riittää vain sekunnin murto-osaksi heijastettu kuvajainen uudesta nopeusrajoituksesta, vaikka vallitsevaan tieympäristöön liittyvää syy-yhteys olisi ollut niin sanotusti täysin epälooginen ja jopa vaarallinen.

Asioiden konteksti toimii syy-yhteyden tavoin. MC.AI (2019) keräämän artikkelin mukaan tekoälyllä on vaikeuksia ymmärtää kontekstia. Ihmisen elekieli ja viestintätyyli ovat aina mukana keskusteluissa. Puhekielessä lauseilla ja painotuksilla on eri merkityksiä riippuen siitä kuka, milloin, missä, miksi ja miten puhuu. Tekoäly ei ymmärrä näitä merkityksiä ollenkaan (MC.AI, 2019). Syy-yhteyden ja kontekstin ymmärtämisen puute on haavoittuvuus, jota voidaan käyttää tekoälyä vastaan.

Hurleyn ja Potterin (2020) mukaan johtajat ovat perinteisesti luottaneet ennen kaikkea havaintoihin, intuitioon ja kokemuksiin omassa päätöksenteossaan. Nyt digitaaliaikakaudella panokset ja mahdolliset seuraukset ovat liian suuret, jotta päätöksentekijä voisi jatkaa edellä kuvatulla tavalla. Nyt päätöksiä tuetaan datalla, analyyseillä ja tekoälyllä. Tiedot ja niiden kautta tehdyt analyysit on sisällytetty johtajien päätössalkkuihin tavalla, joka parantaa heidän kykyään menestyä (Hurley ja Potter, 2020).

Päätöksenteon tukena oleva tekoäly ei vielä pysty argumentoimaan ja perustelemaan omia ajatuksiaan. Se ei pysty antamaan rakentavaa kritiikkiä tai vuorovaikuttamaan osana päätöksentekoprosessia. Tämä tulee esille organisatorisissa ja hierarkkisissa järjestöissä tai yrityksissä. Näihin vinoutumiin kiinnittää huomiota muun muassa Ollila. Ollilan (2019) mukaan organisatoriset vinoutumat tulee huomioida. Ne voi syntyä organisaation kulttuurista, johtajien näkemyksistä tai strategisesta fokuksesta. Organisatorinen vinouma vaikuttaa datan valintaan ja käyttöön. Ne levittäytyvät ydintoimintoihin, jolloin ovat systemaattisia ja ulottuvat laajemmalle kuin yksilöllinen vinouma. Organisatorisen vinouman lähteenä on ylimmän johdon suoraan tai rivien välistä tulkitseminen, datalähteiden painottaminen tai datalähteiden priorisointi. Tämä voi tapahtua

esimerkiksi niiden historiallisen syyn tai helpon saatavuuden takia (Ollila, 2019). Tämä voi johtaa siihen, että tekoälyn ominaisuuksia yliarvostetaan. Băjenescu (2018) mukaan tässä piilee vaara, sillä tekoälyjärjestelmät eivät ymmärrä suorittamiaan tehtäviä. Se vain luottaa koulutusmateriaaliinsa ja saamaansa opetukseen. Tekoäly on vielä kaukana erehtymättömästä (Băjenescu, 2018).

Datalähteiden painottaminen tai datalähteiden priorisointi voivat lisätä tekoälyn haavoittuvuutta. Chiversin (2019) mukaan nämä liittyvät toimenpiteiden rajaamiseen. Rajaamattomana tekoäly voi tehdä asioita toisin kuin sen suunnittelija on halunnut (Chivers, 2019). Toisin sanoen tekoäly ei ymmärrä, että sitä voidaan huijata. Patel ym. (2019) mukaan pahantahtoiselle toimijalle riittää usein se, että tekoälyn haavoittuvuuksien kautta järjestelmä saadaan vain tuottamaan vääriä tuloksia. Tekoäly ei ymmärrä, että sen tuotos on väärää. Lisäksi se ei kiinnitä siihen mitään huomiota. Kataja (2020) antaa Yle.fi artikkelissa aiheeseen sopivan esimerkin. Artikkelin mukaan eräs taiteilija veti perässään kärryä pitkin Berliinin katuja. Kärryssä oli 99 matkapuhelinta. Puhelinten samasta paikasta antamat signaalit saivat Google Maps-karttasovelluksen päättelemään, että alueella oli liikenneruuhka. Karttasovellus muutti kadunosat vihreästä punaisiksi ruuhkan merkiksi ja sovellus suositteli autoilijoita välttämään aluetta (Kataja, 2020).

7 YHDISTELMÄ

Tekoäly ei ole mikään uusi keksintö. Siitä on aloitettu puhumaan aina toisen maailmansodan jälkimaininkien ja 1950-luvun Darthmouthin kesäseminaarin jälkeen. Honkelan (2017) mukaan tekoälyn historia on käytännössä yhtä vanha kuin tietokoneiden. Tekoälyä aloitettiin suunnittelemaan ja kehittämään heti, kun tutkijat saivat käyttöönsä siihen soveltuvat laitteet. Tämän jälkeen tekoälyn erilaiset osa-alueet ovat pysyneet samoina viimeiset viisikymmentä vuotta. Lähtökohta tekoälyn kehittämiselle on ollut se, mitä ihminen osaa ja mitä ihmisen kohdalla pidetään älykkäänä toimintana. Nämä älykkäät toiminnat liittyvät menestyksellisellä tavalla toimimiseen ja elämässä selviämiseen. Tekoälyn kehittämisessä pyritään inhimillisten kykyjen esimerkiksi päättelyn, ongelmaratkaisun ja puhutun tai kirjoitetun kielen ohjelmoimiseen tietokoneelle (Honkela, 2017).

Kirjallisuudessa, scifi-elokuvissa ja erityisesti myyntimiesten keskuudessa tekoäly on ollut aina jotain erityistä. Jotain sellaista mikä mullistaa maailmaa. Hiltusen, E ja Hiltusen, K (2014) mukaan tekoälyllä ja ihmisiä viisaammilla koneilla on leikitelty lukuisia kertoja scifituotannossa. Villeimmissä ja dystopisimmissa visioissa superälykkäät koneet ja robotit ottavat vallan ihmiskunnasta. Lopuksi ne orjuuttavat tai tuhoavat rotumme. Vaikka tämä ajatus onkin elänyt pitkään scifissä, se ei ole sitä vielä todellisuudessa (Hiltunen E & Hiltunen K, 2014).

Nämä Hiltusten esille tuomat villeimmät ajatukset eivät ole vain tuulesta temmattuja. Tekoälyyn on kohdistunut Järvisen (2018) mukaan aina isoja odotuksia. Jo 1960-luvulla tekoälyn tuli jättää kielenkääntäjät työttömiksi. Pian kuitenkin huomattiin, että kääntäminen ei ole helppoa. Sanoja ei vain vaihdeta toisiksi. Tietokoneiden prosessitehon kasvu 1980-luvulla herätti ajatukset uudestaan. Suurista odotuksista huolimatta tulokset jäivät jälleen kerran niukoiksi. Tekoälytutkijat alkoivat puhumaan tekoälytalvesta. Yksi merkittävä askel saavutettiin 1997, kun IBM:n Deep Blue tietokone voitti shakin maailmanmestarin Garri Kasparovin. Tuolloin tekoälyä ei ylistetty, sillä shakkia pidettiin strategisena laskemisena. Vuoden 2010-tienoilla internetin yleistyessä puheentunnistus ja kielenkääntäminen nousi esille. Toiminta oli mahdollista, koska useat puhutut lauseet olivat jo esiintyneet internetissä. Analysoimalla lauseita ja peräkkäisten sanojen todennäköisyyksiä, tunnistuksen tarkkuus nousi uudelle tasolle (Järvinen, 2018).

Tutkijoiden keskuudessa tekoälyllä on oma paikkansa. Välillä kiinnostus on ollut intensiivisempää ja välillä taas laimeampaa. Honkelan (2017) mukaan tekoälyn tutkimusta ja kehitystyötä on tehty useita vuosikymmeniä. Monenlaisia osittaisia läpimurtoja on saavutettu, mutta tekoälyllä ei ole koskaan päästy lähelläkään ihmisen tasoista älykkyyttä. Voidaankin sanoa, että tekoälyn kehityksen haasteiden vuoksi alaa koskeva kiinnostus on aaltoilevaa. Ihminen on ollut älylliseltä kapasiteetiltaan liian ylivoimainen tietokoneisiin verrattaessa. Erityisesti ihminen on ollut ylivertainen siinä, miten maailmaa hahmotetaan kokonaisuutena. Tästä käytetään termiä maailmantieto. Maailmantiedolla tarkoitetaan sitä, miten ihminen pystyy hahmottamaan kaikkea maailmassa olevaa tietoa. Eritoten hahmottamaan siellä esiintyvien asioiden välisiä suhteita ja syy-yhteyksiä. Kun tullaan nykypäivään ihmisten ja tietokoneiden välinen kilpailu on

tasoittunut. Tekoälyjärjestelmät ovat alkaneet haastaa ihmistä yhä erilaisimmilla osa-alueilla (Honkela, 2017).

Honkelan (2017) ja tekoälyohjelman loppuraportin (2019) mukaan nykyisten tietokoneiden laskentatehon, -kapasiteetin sekä muistin kasvu, ovat mahdollistaneet tekoälyn kehittymisen. Nämä ovat mahdollistaneet valtavien datamäärien käsittelyn. Tarkoman (2017) mukaan juuri digitalisaatio on tuonut valtavan määrän tietoa ja tuotteita kaikkien ulottuville maailmanlaajuisessa mittakaavassa. Nyt tätä raakadataa voidaan käsitellä, säilöä ja jalostaa. Hän lisää, että digitalisaatio liittyy olennaisesti yhteyksien määrään, yhteyksien kautta saatavaan dataan sekä jalostettuun dataan eli tietoon (Tarkoma, 2017).

Tarkoman (2017) mukaan tämä datan keruu ja hyödyntäminen tietona kuvaavat hyvin viimeistä kymmentä vuotta. Viimeisen kymmenen vuoden aikana on kehitetty kyvykkyyksiä suunnattomien datamäärien laajamittaiseen käsittelyyn. Tästä käytetään nimitystä big datan louhiminen. Suuri datamäärä ei yksin johda älykkääseen toimintaan. Tämän lisäksi tarvitaan päätöksentekoa tukevia algoritmeja. Big data -ratkaisut ja algoritmit ovat luoneet pohjan tehokkaiden ja laajamittaisten tekoälyratkaisujen kehittämiselle ja käyttöönotolle (Tarkoma, 2017). Honkela (2017) jatkaa, että näiden erittäin laajojen aineistojen syöttäminen tietokoneelle on mahdollistanut tekoälyn kehittämisen ja eräänlaisen keinotekoisen kokemuksen luomisen. Kun tietokone käy läpi laajoja aineistoja, se muodostaa aineistosta malleja. Tämä prosessi muistuttaa ihmisen oppimista. Ihmisten oppimista matkimalla, koneet ovat kehittyneet nopeasti. Erityisesti verrattaessa tätä ohjelmointiin, jossa kirjoitetaan niiden muistiin tietoja ja sääntöjä. Tämän tyyppinen tiedostamattoman tiedon hyödyntäminen on ratkaisevaa ihmisen kykyä matkivien tai jopa ylittävien järjestelmien luomisessa (Honkela, 2017).

Tekoälystä on havaittu myös negatiivisia puolia. Tekoälytutkijat ovat alkaneet kiinnittämään huomiota tekoälyn reaalimaailman uhkiin. Tekoälyohjelman loppuraportin (2019) mukaan tekoälyllä on merkitystä jopa maailman kokonaisturvallisuuden näkökannalta. Tämä väite perustuu siihen, että kaikki suurvallat tavoittelevat tekoälyn edelläkävijän roolia. Taloudellinen kilpailuetu on vain osa kokonaisuutta. Toinen peruste tulee siitä, että yhteiskunnan digitalisaation myötä olemme yhä riippuvaisempia digitaalisesta teknologiasta ja tekoälypohjaisista järjestelmistä. Vaikutusta lisää näiden järjestelmien keskinäisriippuvuus toistensa toiminnasta ja datasta. Sähkökatkot, viestintäverkkojen häiriöt ja digitaalisten järjestelmien toimintahäiriöt voivat rampauttaa tai pysäyttää eri organisaatioiden, sektoreiden tai koko yhteiskunnan toiminnan (Tekoälyohjelman loppuraportti, 2019). Toisaalta raportteja lukiessa ei pidä vain keskittyä uhkiin ja väärinkäyttöihin. Tekoäly on myös mahdollistaja, mainittujen asioiden ja toimintojen, kyber- sekä tietoturvallisen ympäristön ylläpitäjä. Hurleyn ja Potterin (2020) mukaan meidän on lähestyttävä tekoälyn käyttöä käytännönläheisesti ja realistisesti. Hypetystä on hillittävä. On tärkeää keskustella tekoälystä asiallisessa kontekstissa, koska se ei ole ainoa oikea ratkaisu jokaiseen ongelmaan. On haasteita, joiden ratkaisuun sopivat paremmin perinteisemmät menetelmät (Hurley & Potter, 2020).

Edellä kuvattujen uhkien dramatisointi ja pelkojen esille nostaminen on usein seurausta tietämättömyydestä tai tiettyyn yksittäiseen kokonaisuuteen keskittymisestä. Tekoälyyn liittyy paljon uusia mahdollisuuksia, joiden mukana

tulee ymmärrettävästi myös negatiivisia vaikutuksia. On tärkeää tunnistaa tekoälyyn liittyvät uhat ja ymmärtää sen haavoittuvuudet. Honkelan (2017) mukaan tekoälyllä ennakoidaan olevan vaikutuksia arkisissa asioissa ja lähes kaikkien ihmisten elämässä tulevina vuosikymmeninä. On tärkeää ymmärtää mistä on kysymys. Tietämättömyys voi johtaa ahdistukseen ja pelkoihin tulevasta kehityksestä. Myös näitä huolia ja riskejä saatetaan paisutella. Tällöin esimerkiksi tekoälyn monet myönteiset asiat saattavat jäädä huomaamatta. Jos pelkäämme ja ahdistumme, saatamme jättää huomioimatta tekoälyn monet konkreettiset riskit (Honkela, 2017).

Honkelan (2017) mukaan monilla aloilla tekoälyratkaisut ovat arkipäivää. Mediassa korostetaan niin sen uhkia, kuin mahdollisuuksia (Honkela, 2017). Honkelan kanssa samaa mieltä ovat Page, Bain ja Mukhlish (2018). Heidän konferenssijulkaisunsa mukaan tekoälyjärjestelmien eduista huolimatta nyt on tunnistettava myös riskejä. Nyt suunnitellut tekoälyjärjestelmät toimivat tavoilla, joiden riskejä on vaikea ennustaa. Toisaalta meidän tulee olla varovaisia, ettemme esitä perustelemattomia väitteitä. Perustelemattomat väitteet johtavat vain yleiseen epävarmuuteen. Perustellut väitteet mahdollisista riskeistä on kuitenkin esitettävä. Tekoälyn epäonnistuessa pistokkeen vetäminen seinästä tai manuaalinen haltuunotto ei ole todennäköisesti toimivin ratkaisu. On tehtävä tutkimusta näiden mahdollisten tulevaisuuden riskien käsittelemiseksi (Page, Bain & Mukhlish, 2018).

Tekoälyn riskejä ovat muun muassa hyökkäykset näitä järjestelmiä vastaan. Tekoälyä kohtaan suuntautuvia hyökkäyksiä on tunnistettu useita, mutta ei varmasti vielä kaikkia. Luottamuksellisuus-, eheys-, saatavuus- ja replikointihyökkäyksien lisäksi on tulossa joukko muita. Tämä väite perustuu siihen, että vahvuuksistakin voidaan kaivaa esiin heikkouksia. Patel ym. (2019) mukaan suurin osa tähän päivään mennessä tekoälyjärjestelmistä löydetyistä hyökkäysvektoreista on vain pieni osa kaikista mahdollisista tavoista. Esimerkiksi tekoälyn voidaan kuvitella noudattavan sääntöjä kuuliaisesti. Näin tekevät esimerkiksi autonomiset autot. Tämä on tekoälyn yksi haavoittuvuus. Se on mahdollinen hyökkäysvektori muiden joukossa. Järvisen (2018) mukaan ohjeita orjallisesti noudattavat autot aiheuttavat vahinkoa muille tai omille matkustajilleen, jos esimerkiksi liikennemerkkejä on väärennetty. Ihmiskuljettaja osaa arvioida, milloin sääntöjä pitää rikkoa tai milloin sääntöjä on vääristelty. Tekoäly ei tähän pysty. Tulevaisuuden robottiautoissa tietoturvariskit tulevat olemaan moninkertaisia. Nämä autonomiset autot tulevat olemaan tietoturvan painajainen (Järvinen, 2018). Ghafarian ja Sardarin (2020) mukaan autonomisen ajoneuvon ja sen taustajärjestelmän välinen viestintä on yksi turvallisuusongelma. Näihin kuuluvat viestintä auton navigoinnin kanssa tai muu viestintä, sekä kauko-ohjaus. Esimerkiksi haittaohjelmahyökkäys viestintäjärjestelmään voi aiheuttaa vääristymiä autonomisen auton navigoinnissa. Tällöin on olemassa vaara, että ajoneuvo ajaa väärällä nopeudella sijaintiinsa nähden (Ghafarian & Sardari, 2020).

Uuden tuottamisen kilpajuoksu on taloudellisesti perusteltua, mutta näiden uusien ratkaisujen haavoittuvuusuhkia pitäisi myös miettiä etukäteen. Halusen (2018b) mukaan olisi hyvä miettiä, miten tekoälyjärjestelmä tulisi suojata sen erilaisia haavoittuvuuksia vastaan. Nyt olemme nähneet tekoälyn hyödyistä jäävuoren huipun. Samalla olemme nähneet tekoälyn haavoittuvuuksistakin

vain samanlaisen huipun. Pinnan alla olevan osa on vielä tutkimatonta. Tekoälyn tulevat hyödyt ja haavoittuvuudet tuovat mukanaan uusia ja arvaamattomia kehityskulkuja (Halunen, 2018b). Halusen kanssa samaa mieltä on Patel ym. (2019). Heidän laatiman raportin mukaan tekoälyyn kohdistuvien hyökkäysten vastainen toiminta on edennyt käsi kädessä hyökkäysten tutkimuksen kanssa. Luonnollisesti akateemisessa ympäristössä on tutkittu paljon enemmän sitä, kuinka puolustetaan päivittäin hyökkäyksen kohteena olevia järjestelmiä, kuin miten niihin mahdollisesti hyökättäisiin (Patel ym., 2019).

Tekoälyn kybertoimintaympäristön haavoittuvuuksissa on monta eri kohdetta. Oppimisen, opettamisen, koulutusmateriaalin, algoritmien, koodi- ja ohjelmistotason, tunnistustekniikoiden, ymmärryksen sekä laitetason haavoittuvuudet olisi minimoitava. Aivan ensiksi nämä haavoittuvuudet olisi kuitenkin tunnistettava ja ymmärrettävä. Tämän jälkeen voimme vasta suojautua niiden kautta tapahtuvalta pahantahtoiselta toiminnalta. Tämä on helpommin sanottu kuin tehty. Esimerkiksi STT:n (2019) artikkelin mukaan tekoäly ei kerro käyttäjälleen, miten se on tiettyyn omaan ratkaisuunsa päätynyt. Minkä haavoittuvuuden tällöin minimoimme? STT:n (2019) artikkeli esittää ratkaisuksi, että tekoälyjärjestelmät on rakennettava heti mahdollisimman turvallisiksi. Ajatus on oikea, mutta toteuttaminen onkin taas eri asia. On esimerkiksi mietittävä yksittäisiä haavoittuvuuksia, sekä järjestelmäkokonaisuuden kyberturvallisuutta. Tilaa pitää myös jättää tekoälyratkaisujen kehittymiselle. Lehdon (2019) mukaan useiden yksittäisten tekoälyratkaisujen haasteena on asioiden sirpaleisuus ja vajaavaisuus. Toiminnan kehittymispotentiaalin huomioiminen edellyttää integroitumisen mahdollistamista ja integraatiojärjestelmiä. Lehdon (2019) mukaan integroiduissa järjestelmissä on omat ulkoiset ja sisäiset uhat, jotka tulee huomioida. Uusi integrointi aiheuttaa kompleksisuuden kasvua alkuperäiseen verrattuna. Lopulta yksittäisistä tekoälyratkaisuista koostuviin kokonaisuuksiin on rakennettava kokonaisvaltainen kyberturvallisuusjärjestelmä (Lehto, 2019).

Ihmismäistä tekoälyä ei ole vielä saavutettu. Tulevaisuudessa nykyinen kapea tekoäly on kehittymässä kohti tulevaisuuden yleistä tekoälyä. Ailiston ym. (2018) mukaan tekoälyn nopea kehittyminen on tuonut mukanaan suuria toiveita, mutta myös pelottavia uhkakuvia tekoälyn tunnistamattomien ominaisuuksien suhteen. Vaikka emme ole vielä saavuttaneet yleistä tekoälyä, on hyvä tiedostaa tekoälyn ominaisuuksien kehittyminen. Tämä voi olla tietyillä sovellusalueilla hyvinkin nopeaa (Ailisto ym., 2018). Ailiston ym. tapaan myös Järvinen nostaa esille yleisen tekoälyn. Järvisen (2018) mukaan tekoälyn suurin kysymys on: Voidaanko kapeasta tekoälystä edetä yleiseen tekoälyyn? Yleisessä tekoälyssä ohjelma tai kone oppisi mitä tahansa logiikkaa ja päättelyä vaativia tehtäviä. Tällöin käyttömahdollisuuksia olisi rajattomasti. Tämän tason saavuttaminen ei ole kaukana. Jos tietokoneet kehittyvät nykyistä vauhtia, yleinen tekoäly ei jää edes alansa viimeiseksi saavutukseksi. Riittävän edistynyt yleinen tekoäly ohittaa ihmisen ja alkaa kehittämään itseään. Itseään kehittävä tekoäly johtaa exponentiaaliseen kierteeseen, jossa tekoäly lopulta syrjäyttää ihmisen. Tekniikka karkaa tasolle, jota kutsutaan singulariteetiksi tai supertekoälyksi (Järvinen, 2018). Supertekoälystä kirjoittaa muun muassa Bostrom (2014) kirjassaan: Superintelligence: Paths, Dangers, Strategies.

Kaikki tekoälyä pohtineet eivät allekirjoita kehityksen tai hypetyksen jarruttelua. Ailiston ym. (2018) ja Järvisen (2018) nostamista uhkakuvista huolimatta Siukosen ja Neittaanmäen (2019) mukaan tekoälyn kehitystä ei voi, eikä ole tarpeenkaan pysäyttää. On tärkeää pysyä tekoälykehityksen mukana ja oppia samalla. Tutkimusten mukaan yritykset ympäri maailmaa pitävät tekoälyä strategisesti tärkeänä johtamisen, liiketalouden ja tuotannon kannalta (Siukonen & Neittaanmäki, 2019).

8 JOHTOPÄÄTÖKSET JA POHDINTA

Tekoäly on tehokas työkalu, mutta sillä on haavoittuvuuksia. Sitä kuinka paljon ja kuinka vahingollisia haavoittuvuuksia tekoälyjärjestelmät sisältävät ei vielä tiedosteta. Kirjallisuuskatsauksen pohjalta tekoälyhaavoittuvuuksien syvällinen tutkimus on vasta alussa. Tieteellistä teoriapohjaa ”oikean vastauksen” antamiseen rakennetaan vasta. Tekoälyn käyttö on kiihtyvä luonnonvara, jota on ymmärrettävä. Tutkimuksen perusteella osa tekoälytutkijoista on nostanut esille haavoittuvuuksien ymmärtämisen puutteen. Näiden argumenttien perusteella tekoälyn haavoittuvuuksien tutkimukselle on ollut ja on edelleenkin tieteellinen tarve. Erilaisten näkökulmien esittäminen ja kontekstiin liittyvien tutkimusten julkaiseminen vauhdittaa keskustelua aihealueesta. Tutkimuksen tuloksena olevat haavoittuvuudet tulevat koskettamaan huonoimmassa skenaariossa kaikkia tulevia tekoälyn loppukäyttäjiä. Ettei näin kävisi, on tämän aiheen laaja-alainen tieteellinen tutkiminen merkityksellistä.

Tieteellisesti merkityksellinen tutkimus ja sen tulokset on kyettävä osoittamaan luotettaviksi. Tämän tutkimuksen luotettavuutta korostettiin lähteiden synteesillä ja tutkimusaineistolla. Tutkimuksen aineiston analyysiprosessi jaettiin kolmivaiheiseksi. Tutkimusaineisto pidettiin laajahkona, jotta tutkimuksen kannalta relevanttisimmat tiedot löydettiin useammasta kuin yhdestä lähteestä. Keskeisimmät lähteet myös korreloituivat keskenään erityisesti konferenssijulkaisujen ja elektronisten kausijulkaisujen kesken. Lisäksi näiden julkaisujen taustalta löytyi samaa lähdeaineistoa, kuin mitä itse tutkimuksessa käytettiin. Lähteiden synteesillä ja aineiston analyysiprosessin noudattamisella saavutettiin tutkimustulosten tieteellinen merkityksellisyys, sekä tutkimuksen reliabiliteetti ja validiteetti. Tutkimuksen luotettavuudessa painopisteeksi muodostui oikeiden ja rajattujen asioiden tutkiminen, sekä tulosten pysyvyys ja konkretia.

Tutkimukselle asetettuun päätutkimuskysymykseen saatiin vastaus. Tekoälyllä on haavoittuvuuksia kybertoimintaympäristön näkökulmasta. Kybertoimintaympäristössä tekoälyn haavoittuvuudet liittyvät sen tekniseen toteutukseen, perustarpeisiin ja toimintatapaan. Tekniseen toteutukseen liittyvät tekoälyn koodi-, ohjelmisto- ja laitetason haavoittuvuudet. Perustarpeisiin liittyvät oppimisprosessin haavoittuvuudet. Näitä oppimisprosessin haavoittuvuuksia löytyy tekoälyn oppimisesta, opettamisesta ja tähän käytettävästä koulutusmateriaalista. Tekoälyn toimintatapaan liittyviä haavoittuvuuksia löytyy tunnistustekniikoista, algoritmeista ja ymmärryksestä. Lisäksi on hyvä ymmärtää, että kaikki edellä mainitut haavoittuvuudet löytyvät kuin koonnoksena monimutkaisimmista tekoälyjärjestelmäkokonaisuuksista. On vain valittava, mitä haavoittuvuutta haluaa hyödyntää.

Tekoälyn kehittämisen yksi tärkeimmistä vaiheista on opettaminen ja oppiminen. Tekoälyä opetetaan esimerkiksi ohjatusti, ohjaamattomasti tai vahvistusoppimisen kautta. Tekoälyn opettamistekniikat riippuvat ratkaistavasta tilanteesta, opetukseen käytettävissä olevasta tiedosta, koulutusmateriaalista ja tekoälyn loppukäytöstä. Lähdekirjallisuuden ja tutkimustulosten perusteella näihin tekoälyn ensiaskeliin liittyy haavoittuvuuksia. Opetusvaiheessa on riski, että tekoälylle ei identifioida tai huomioida tosielämässä eteen tulevia tilanteita. Jos tätä

ei tehdä, on mahdollista huijata tekoälyjärjestelmää. Huijaus voi mennä läpi tekoälylle, mutta ei onnistu saman opetuksen käyneelle ihmiselle. Tämä johtuu esimerkiksi siitä, että tekoäly tekee juuri niin kuin ihminen on sen opettanut. Näitä monimutkaisia tilanteita voidaan kompensoida käyttämällä tekoälyn opettamiseen hienostuneita oppimismalleja. Toisaalta tekoälylle on vaikea opettaa kaikkia mahdollisia vastaantulevia tilanteita. Oppimiseen jää aina aukkoja, joita ihminen oppisi täyttämään syy-yhteyden omaksumisen kautta. Tekoäly ei omaksu asioita. Edellisten lisäksi pahantahtoinen toimija voi päästä myös peukaloimaan itse opetusprosessia. Tutkimustulosten mukaan peukaloitu opetusprosessi mahdollistaa tekoälyn toistamaan sille opetettuja vääristyneitä toimintamalleja. Lisäksi tekoäly voi oppia myös vääriä asioita kaikesta huolimatta. Esimerkiksi juuri niitä ”vääriä sanoja”, joita itse on välttänyt käyttämästä lasten kuullen. Sitten nämä väärät sanat kuuluvat autoradiosta, kun olet ulkona tankkaamassa.

Tekoälyn oppiminen vaatii valtavia määriä koulutusmateriaalia. Lähdeaineiston perusteella koulutusmateriaalia tarvitaan itse tehtävästä ja suoritettavien tehtävien ympäriltä. Tekoälyn tarvitsemaan suureen tietomäärään liittyy haavoittuvuuksia. Tutkimustulosten perusteella yksi tekoälyn keskeinen haavoittuvuus on koulutusmateriaalin peukalointi tai vääristäminen. Tulos korreloi usean eri lähdeteoksen ja tekoälytutkijan näkemyksen kanssa. Koulutusmateriaalihaavoittuvuuden yksi vahingoittuvimmista vaiheista on opetusvaihe. Tämän vaiheen aikana tekoäly oppii tai sitä opetetaan koulutusmateriaalin perusteella. Jos pahantahtoinen toimija pääsee käsiksi tähän koulutusmateriaaliin, tekoälyn opettaminen vaarantuu. Tekoäly ei tule toimimaan, niin kuin sen haluttaisiin toimivan. Usean lähteen mukaan koulutusmateriaalia voi vääristää tekoälyn tehtävään liittyen epätarkoituksenmukaisemmaksi tai peukaloida sitä vain omien tarpeiden mukaiseksi. Vaikka tekoälyn koulutusmateriaalia vääristetään vain hieman ja mahdollisimman huomaamattomasti, tämä mahdollistaa halutun lopputuloksen ennen pitkää. On myös huomioitava, että tekoälyn kouluttamiseen tarvittava tieto itsessään voi osoittautua suojeltavaksi arkaluontoisine tietoineen tai liikesalaisuuksineen. Tutkimustulosten perusteella koulutusmateriaaliksi tarkoitettu keskitetty tietovarasto on myös tekoälyn haavoittuvuus. Kerätty arkaluontoinen tieto houkuttelee pahantahtoisia toimijoita.

Tutkimuksen taustakirjallisuus korostaa algoritmien osuutta tekoälyjärjestelmissä. Algoritmeilla pyritään mallintamaan ihmismäistä ajattelua ohjelmallisin keinoin. Algoritmi antaa tekoälylle seikkaperäisiä ohjeita tai selvityksiä, miten sen on suoritettava jokin tehtävä tai prosessi. Tekoälyjärjestelmät koostuvat usein useammasta algoritmista. Tutkimustulosten perusteella nämä tekoälyn algoritmit sisältävät haavoittuvuuksia, jotka vaikuttavat tekoälyn toimintaan negatiivisesti. Tutkimustulosten perusteella algoritmiin liittyvät haavoittuvuudet ovat tekoälykontekstissa välillisiä. Tutkitut haavoittuvuudet eivät liity suoraan algoritmeihin, vaan esimerkiksi niille syötettävään informaatioon. Myös lähdekirjallisuudessa ei käsitellä suoraan itse algoritmien haavoittuvuuksia. Tämä vääristynyt informaatio saa algoritmin tekemään ei toivottuja ratkaisuja. Ei toivottuja ratkaisuja ovat muun muassa väärän luokituksen aiheuttaminen ja oppivan algoritmin manipulointi. Väärä luokitus saadaan esimerkiksi sekoittamalla algoritmin kyky verrata oikeaa sisältöä väärään. Tämä on mahdollista, koska algoritmeilla pyritään yhä enemmän ihmismäiseen loogiseen päätöksentekoon.

Tämä tavoite tuottaa lisähaasteita. Ihmismäisessä päätöksenteossa ei voida sanoa, miksi tai missä kohtaa tekoäly teki huonon ratkaisun. Tekoälyn algoritmien haavoittuvuuksissa täytyy huomioida lisäksi se, että tekijöiden oma ajatusmaailma saattaa siirtyä osaksi niiden toimintaperiaatteita. Tämä voi tapahtua joko tiedostamatta tai tarkoituksella.

Tekoäly koostuu koodista ja ohjelmistoista. Näissä koodeissa ja ohjelmistoissa esiintyy heikkouksia ja niin sanottuja bugeja. Bugeja ja heikkouksia hyödyntämällä voidaan toteuttaa esimerkiksi kyberhyökkäyksiä. Lähdeaineiston mukaan hyökkäyksissä voidaan vääristää tekoälyyn liittyvää koulutusmateriaalia tai varastaa sen käyttämiä arkaluontoisia tietoja. Tämän takia tutkimuksessa mainitaan tekoälyn käyttämän koodin ja ohjelmistojen olevan yksi sen haavoittuvuuksista. Koodiriveihin voi päästä käsiksi esimerkiksi internetin kautta suoraan tai kiertäen. Tämä on mahdollista, jos tekoälyn ohjelmakoodia on julkaistu esimerkiksi internetissä avoimena. Näiden koodi- ja ohjelmistohaavoittuvuuksien kautta pahantahtoiset toimijat voivat päästä järjestelmään, kuten missä tahansa tietokonesovelluksessa tai -verkossa. Edellisten lisäksi on hyvä muistaa, että tietojärjestelmäkokonaisuuksien sisältämien ohjelmistojen määrä ja monimutkaisuus tuottaa loppukäyttäjille haasteita. Tämä näkyy esimerkiksi laitteiden päivitysten ajantasaisina pitämisessä. On mahdollista, että tekoälyn käyttämien päivittämättömien ohjelmistojen tai koodivirheiden haavoittuvuuksia hyödynnetään pahantahtoisessa toiminnassa.

Tekoäly pohjautuu fyysisiin laiteratkaisuihin, kuten mikä tahansa tietojärjestelmä. Laite- ja komponenttitasoilla olevilla fyysisillä laitteilla on olemassa olevia haavoittuvuuksia. Jos kykenee hallinnoimaan tekoälyn käyttämiä laitteistoja, aiheuttaa se uhkan koko järjestelmän toiminnalle. Tutkimuksen lähteiden perusteella tätä tekoälyn haavoittuvuutta ei tunnisteta laajasti. Vain muutamassa lähdeteoksessa tekoälyn toimintatavan vääristäminen on ulotettu laite- ja komponenttitasoille. Monissa lähdeteoksissa painotetaan ohjelmistohaavoittuvuutta (software), eikä tunnisteta laitteistotasoa (hardware). Laitetason haavoittuvuuksilla voi eskaloida tekoälyjärjestelmiin ongelmia, jotka voivat johtaa esimerkiksi tekoälyä käyttävien koneiden hallinnan menettämiseen. Hallinnan menettäminen voidaan toteuttaa esimerkiksi laite- ja komponenttitasoille tehdyillä takaporteilla. Ne voivat olla myös ylimääräisiä komponentteja, jotka muuttavat tekoälyn toiminnallisuutta.

Tekoälyn tunnistustoiminnassa on haavoittuvuuksia, jotka voivat johtaa toiminnallisiin virheisiin esimerkiksi päätöksentekotilanteissa. Pahantahtoinen tekijä voi harhauttaa tekoälyä esimerkiksi hahmontunnistuksessa. Tämä tapahtuu vaikuttamalla fyysisen reaalimaailman esineisiin sekä digitaalisen maailman ominaisuuksiin liittyvillä toimilla. Lähdeaineiston ja tutkimustulosten perusteella tunnistustekniikoiden vääristymät ohjaavat tekoälyn luokittelemaan asioita täysin toiseksi mitä ne oikeasti ovat. Tekoälyn tekemät tunnistusvirheet voivat tuntua ihmisistä käsittämättömiltä. Toisaalta ne voivat olla ihmissilmälle täysin näkymättömiä. Automatisoiduissa ympäristöissä tekoälyn tunnistustekniikoiden haavoittuvuudet muodostavat todellisia uhkia kaikille sen ympärillä oleville.

Tekoälyllä on haasteita ymmärtää asioiden kontekstia. Ihmisten välisessä toiminnassa erilaiset viestintätyylit ja asiayhteydet ovat aina läsnä. Tekoäly ei

ymmärrä näiden merkityksiä omassa toiminnassaan. Tutkimustulosten perusteella tämä asioiden syy-yhteyden ymmärtämisen puute on tekoälyn yksi haavoittuvuus. Tämä todetaan myös tutkimuksen useassa eri lähdeteoksessa. Ongelmallista on, että tätäkin haavoittuvuutta voidaan käyttää tekoälyä vastaan negatiivisessa mielessä. Tekoäly löytää helposti sille opetetuille asioille vastaavuussuhteita, mutta ei ymmärrä niiden syy-yhteyttä. Tekoäly esimerkiksi löytää koulutusmateriaalissaan olleita hahmoja vastaavia kuvioita sille annetusta aineistoista, mutta se ei ymmärrä missä kaikissa eri yhteyksissä kuviota voi esiintyä. Tämä on ongelmallista erityisesti tunnistamisen jälkeen tapahtuvassa päätöksentekoprosessissa. Tekoäly ei todistele tai argumentoi, miksi ja miten se teki kyseisen päätöksen. Sillä ei ole kykyä vuorovaikutukseen.

Tekoälyn haavoittuvuuksien kautta tapahtuvan pahantahtoisen toiminnan kiinnostuksen kohteita on lukematon määrä. Ne voivat olla esimerkiksi sisällönsuodattimen ohittaminen, taloudellinen aspekti, tekoälyn luottamuksen myrkyttäminen tai tekoälyn antaman tiedon vääristäminen. Lisäksi tavoitteena voi olla tekoälyn käyttämään kriittiseen tietoon pääseminen, kiusanteko tai tekoälyn kokoaman, suodattaman ja seuloman tiedon varastaminen. Näihin tekoihin käytetään tekoälyyn perustuvien järjestelmien vastaisia hyökkäyksiä. Tekoälyä vastaan tapahtuvista hyökkäyksistä kirjoitetaan monessa lähdeteoksessa. Usein hyökkäysten tarkempaa jakoa ei kuitenkaan tehdä. Tekoälytutkijoiden keskuudessa muutama tieteentekijä on perehtynyt hyökkäysten erilaisiin toteutustapoihin. Tutkimuksen yhtenä tuloksena on, että tekoälyä vastaan voidaan hyökätä luottamuksellisuus-, eheys-, saatavuus- tai replikointihyökkäyksin. Nämä ovat keskeisiä offensiiveja tekoälyä vastaan. Tämä johtopäätös vastaa suoraan tutkimuksen ensimmäiseen apukysymykseen.

Myös toiseen apututkimuskysymykseen saatiin vastaus. Tekoälyn haavoittuvuudet ja kyberturvallisuus linkittyvät toisiinsa. Nyky-yhteiskuntaa ja sen yksittäisiä osia voidaan vahingoittaa edellä mainittujen tekoälyn haavoittuvuuksien kautta. Hyödyntämällä tekoälyn haavoittuvuuksia pahantahtoiset toimijat pystyvät toteuttamaan päämääränsä kybertoimintaympäristössä. Toiminta voi kohdistua itseohjautuviin ja autopilotilla liikkuviin liikennevälineisiin, teollisuusrobotteihin, erilaisiin viestintäjärjestelmiin, rahoitusliikenteeseen tai mihin tahansa automatisoituun tekoälyjärjestelmäkokonaisuuteen. Tekoälyn haavoittuvuuksia hyödyntämällä voidaan sekoittaa nyky-yhteiskunnan lämmön-, veden- sekä sähkönjakelu ja niin edelleen. Haavoittuvuuksien kautta on mahdollista horjuttaa tai romahduttaa haluamansa loppukäyttäjän, organisaation tai palvelun toiminta ja heikentää ihmisten luottamusta näihin.

Tässä tutkimuksessa esille tulleet tekoälyn haavoittuvuudet kybertoimintaympäristössä ovat vain pisaroita meressä, jotka ymmärrämme tekoälyn yhdestä osa-alueesta tänään. Toisaalta kun tarkastelee tämän pro gradun tuloksia, ei ole vaikea kuvitella tämän aihealueen haasteita ja skenaarioita tulevaisuudessa. Tekoäly on nyky-yhteiskuntamme olennainen lisä. Se koskettaa läheisesti modernin yhteiskuntamme eri osia. Näillä tutkimustuloksilla ja johtopäätöksillä on merkitystä palveluyhteiskuntamme käyttämän tekoälyn paremmassa ymmärtämisessä. Saadut tutkimustulokset lisäävät tekoälyn käytettävyyttä tulevaisuudessa, sillä ymmärrämme paremmin sen haavoittuvuuksia. Vaikka tulevat haavoittuvuudet voivat jalostua pahemmiksi kuin tutkimustulokset nyt esittävät,

olemme yhden askeleen valmiimpia kohtaamaan ne. Tekoäly tarjoaa tulevaisuudessa varmasti paljon uusia hyviä mahdollisuuksia. Sen takia meidän on otettava se vakavasti. On ymmärrettävä tekoälyn vahvuudet ja haavoittuvuudet tai muuten saatamme joutua itse tukemaan sitä.

9 JATKOTUTKIMUKSET

Tekoälyn ympärillä moni asia on uutta ja tutkimatonta. Kaikkia perusteitakaan ei ole vielä selvitetty. Tekoäly-ympäristöjen osalta myös muut kuin kybertoimintaympäristö ovat täynnä kartoittamatonta maaperää. Tämän lisäksi tekoälyn kehityskulku on niin huimaa, että voimme lukea uusista tekoälyjärjestelmistä ja -innovaatioista lähes päivittäin. Näiden johdosta tekoälyä kohtaan on valtavasti jatkotutkimuskohteita ja tutkimattomia alueita. Lisäksi tekoälytutkimus tulee nousemaan osaksi muita tieteitä ja uudistamaan tutkimuksen tekemistä esimerkiksi lähdeaineiston käsittelyn ”automatisoinnin” osalta.

Tekoälyllä on haavoittuvuuksia. Nämä haavoittuvuudet tulisi saada julkisesti saataville ja useampien tutkijoiden, suunnittelijoiden, tekoälyn rakentajien sekä loppukäyttäjien tietoon. Tämä prosessi vaatisi osakseen tutkimusta. Miten haavoittuvuudet paljastetaan, sekä miten ne raportoidaan tai julkaistaan? Myös tekoälytutkijat ovat nostaneet tämän asiakokonaisuuden esille. Brundagen ym. (2018) mukaan tekoälyhaavoittuvuuksien vastuullinen paljastaminen on keskeistä. Vastuullisen paljastamisen lisäksi tekoälyn haavoittuvuuksille on luotava menettelyt luottamuksellisen raportoinnin mahdollistamiseksi. Tekoälyjärjestelmissä havaittuja haavoittuvuuksia ja tietoturvahaavoittuvuuksia on kyettävä julkaisemaan, jotta ne saavuttavat tekoälyn kanssa tekemisessä olevat loppukäyttäjät ja muut toimijat (Brundage ym., 2018). Shevlane ja Dafoe (2020) antavat mielenkiintoisen näkemyksen ylläolevaan Brundagen kannanottoon. Heidän tekemässään konferenssijulkaisussa pohditaan sitä, vähentääkö tekoälytutkimuksen julkaiseminen sen väärinkäyttöä? Konferenssijulkaisussa pohditaan sitä, antaako tutkimuksen julkaisu hyötyä myös hyökkääjille? On mahdollista, että puolustautumista koskeva tieto voidaan muuntaa tehokkaammaksi hyökkäykseksi. Tällöin tosin hyökkääjien pitää omaksua ja soveltaa tutkimustuloksia. Asia ei ole yksinkertaista, sillä tutkimustulosten siirrettävyys puolustuksesta hyökkäykseen tai päinvastoin vaihtelee aina tapauskohtaisesti. Lisäksi asia on riippuvainen useista tekijöistä, kuten kuinka paljon tietoa paljastetaan, minkä kanavan kautta ja miten se esitetään (Shevlane & Dafoe, 2020)?

Systeemianalyysillä on pyritty useissa aiheissa saavuttamaan kokonaisoptimointia. Tätä on tehty muun muassa operaatioanalyysiä ja operaatiotutkimusta mallintamalla. Tavoitteena on ollut saada esille niin sanotusti rivien välissä oleva taso eli metataso. Tämä taso on tutkimusgenressä keskeinen, koska sen avulla tieteellisen tekstin kirjoittaja luo uusia ajatuksia vastaamalla kysymykseen miksi. Tarkoman (2017) mukaan tekoäly on maailmanlaajuinen mahdollisuus lukuisilla eri sektoreilla. Tekoälyyn liittyy myös erilaisia uhkakuvia, jotka vaativat pitkälle meneviä ratkaisuita. Näitä uhkia voitaisiin mahdollisesti estää ja torjua metatasolta alkaen. Metataso mallintaisi ja seuraisi tekoälypohjaisen järjestelmän toimintaa laajamittaisesti. Metataso estäisi tekoälytutkimuksen harhailun ja auttaisi sitä selvittämään itse kontekstiin liittyviä asioita. Tätä metatasoa ei vielä ole, joten se on lähitulevaisuuden mielenkiintoinen tutkimuskohde (Tarkoma, 2017).

LÄHTEET

(a) Artikkeli konferenssijulkaisussa:

Ghafarian, A. & Sardari, S. (2020). An Analysis of Connected Cars Technology and Security. Teoksessa International Conference on Cyber Warfare and Security 2020 vol. XIV (195-203). United Kingdom: Academic Conferences International Limited.

Greiman, V. A. (2020). Artificially Intelligent Systems and Human Rights: A Global Perspective. Teoksessa International Conference on Cyber Warfare and Security 2020 vol. XIV (204-210). United Kingdom: Academic Conferences International Limited.

He, H. Gray, J. Cangelosi, A. Meng, Q. McGinnity, T. M. & Mehnen, J. (2020). The Challenges and Opportunities of Artificial Intelligence for Trustworthy Robots and Autonomous Systems. 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE). United Kingdom: IEEE.

Hurley, J. S. & Potter, D. O. Avoiding the Pitfalls of an Artificial Reality (Is A.I. Real Enough). Teoksessa International Conference on Cyber Warfare and Security vol. XIV, XVII (236-242). United Kingdom: Academic Conferences International Limited.

Page, J., Bain, M. & Mukhlish, F. (2018). The Risks of Low Level Narrow Artificial Intelligence. Proceedings of the 2018 IEEE International Conference on Intelligence and Safety for Robotics. Shenyang, China, August 24-27, 2018.

Shevlane, T. & Dafoe, A. (2020). The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? AIES '20 AAAI/ACM Conference on AI, Ethics, and Society. (173–179). New York, NY, USA.

Vähäkainu, P., Lehto, M. & Kariluoto, A. (2020). IoT –based Adversarial Attack's Effect on Cloud Data Platform Services in a Smart Building Context. Teoksessa International Conference on Cyber Warfare and Security 2020 vol. XVII (457-465). United Kingdom: Academic Conferences International Limited.

(b) Kirja:

Alasuutari, P. (2001). *Laadullinen tutkimus*. (4.uud. painos). Jyväskylä: Gummerus kirjapaino Oy.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chivers, T. (2019). *The AI does not hate you*. Great Britain: Clays Ltd, Elocograf S.p.A.

Fry, H. (2018). *Hello world.* Liettua: Scandbook UAB.

Haikonen, P. O. A. (2017). *Tietoisuus, tekoäly ja robootit.* Tallinna: Printon.

Hiltunen, E. & Hiltunen, K. (2014). *Teknoelämää 2035. Miten teknologia muuttaa tulevaisuuttamme?* Helsinki: Talentum.

Hirsijärvi, S. Remes, P. & Sajavaara, P. (2000). *Tutkija kirjoita.* (6.uud. painos). Vantaa: Kirjayhtymä Oy.

Hirsijärvi, S., Remes, P. & Sajavaara, P. (2004). *Tutki ja kirjoita.* (10. osin uudistettu painos). Jyväskylä: Gummerus Kirjapaino Oy.

Honkela, T. (2017). *Rauhankone. Tekoälytutkijan testamentti.* Helsinki: Tallinna Raamatutrükikoja OÜ

Järvinen, P. (2018). *Kyberuhkia ja somesotaa.* Jyväskylä: Docendo Oy.

Järvinen, P. & Järvinen, A. 2004. *Tutkimustyön metodeista.* Tampere: Tampereen Yliopistopaino Oy.

Jääskeläinen, A. (2019). *Mitä tapahtuu huomenna kun tekoäly poistaa järjettömyydet?* EU: WSOY.

Kananen, J. (2013). *Case-tutkimus opinnäytetyönä.* Suomen Yliopistopaino Oy – Juvenes Print.

Kananen, J. (2014). *Laadullinen tutkimus opinnäytetyönä.* Suomen Yliopistopaino Oy – Juvenes Print.

Kananen, J. (2015). *Opinnöytetyön kirjoittajan opas. Näin kirjoitan opinnäytetyön tai pro gradun alusta loppuun.* Suomen Yliopistopaino Oy – Juvenes Print.

Kananen, J. (2019). *Opinnäytetyön ja pro gradun pikaopas. Avain opinnäytetyön ja pro gradun kirjoittamiseen.* PunaMusta Oy.

Kokkarinen, I. (2003). *Tekoäly, laskettavuus ja logiikka.* Saarijärvi: Gummerus Kirjapaino Oy.

Koskinen, I. Alasuutari, P & Peltonen, T. (2005). *Laadulliset menetelmät kauppatieteissä.* Tampere: Vastapaino.

Laine, M. Bamberg, J. & Jokinen, P. (2007). *Tapaustutkimuksen taito.* Helsinki: Gaudeamus Helsinki University Press.

Laitila, E. (2019). *Vastuullisen tekoälyn ohjelmointi*. Turku: Painosalama Oy.

Lehto, M. (2019). Onko tekoäly turvallinen? Teoksessa Siukonen, T. & Neittaanmäki, P. (2019). *Mitä tulisi tietää tekoälystä (299-309).* Jyväskylä: Docendo

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis.* (2. painos). California: Sage.

Ollila, M-R. (2019). *Tekoälyn etiikka.* Helsinki: Kustannusosakeyhtiö Otava.

Puusa, A. & Juuti, P. toim. (2011). *Menetelmäviidakon raivaajat. Perusteita laadullisen tutkimuslähestymistavan valintaan.* Hansaprint.

Sautoy, M. (2019). *The creativity code. How AI is learning to write, paint and think.* Great Britain: CPI Group (UK) Ltd.

Silverman, D. (1993). *Interpreting qualitative data.* Great Britain: The Cromwell Press Ltd.

Siukonen, T. & Neittaanmäki, P. (2019). *Mitä tulisi tietää tekoälystä.* Jyväskylä: Docendo.

Syrjälä, L, Ahonen, S, Syrjäläinen, E & Saari, S. (1996). *Laadullisen tutkimuksen työtapoja.* Rauma: Kirjapaino Oy West Point.

Tegmark, M. (2018). *Elämä 3.0 Ihmisenä oleminen tekoälyn aikakaudella.* Libris: Terra Gognita.

Tuomi, J. (2007). *Tutki ja lue. Johdatus tieteellisen tekstin ymmärtämiseen.* Jyväskylä: Gummerus Kirjapaino Oy.

Tuomi, J. & Sarajärvi, A. (2009). *Laadullinen tutkimus ja sisällönanalyysi.* (10. uudistettu painos). Vantaa: Hansaprint Oy.

Tähtinen, S. (2005). *Järjestelmäintegraatio.* Jyväskylä: Talentum media OY.

Valli, R. & Aaltola, J. toim. (2015). *Ikkunoita tutkimusmetodeihin 1.* (4.uud. painos). Jyväskylä: PS-kustannus.

Vahvanen, P. (2018). *Kone kaikkivaltias. Kuinka digitalisaatio tuhoaa kaiken meille arvokkaan.* Keuruu: Otavan Kirjapaino Oy.

(c) Elektroninen kirja

Laari, T. (toim.). (2019). #*kyberpuolustus. Kyberkäsikirja Puolustusvoimien henkilöstölle.* Maanpuolustuskorkeakoulu. Sotataidon laitos. Haettu osoitteesta https://www.doria.fi/bitstream/handle/10024/173254/%23kyberpuolustus%20verkko%20%28interaktiivinen%20pdf%29%20%28002%29.pdf?sequence=1&isAllowed=y

(d) Opinnäyte:

Mikkonen, M. (2003). *Järjestelmähallinnasta palvelunhallintaan – Merivoiminen järjestelmähallinnan kehittäminen 2004-2008*. Tutkielma. Maanpuolustuskorkeakoulu, Täydennyskoulutusosasto, Tekniikan Laitos. Viranomaiskäyttö.

(e) Elektronisessa kausijulkaisussa oleva artikkeli:

Băjenescu, T-M. (2018) The risks of artificial intelligence. *Journal of Engineering Science (Chişinău.)* 15(4), 47-56. Haettu osoitteesta https://jes.utm.md/wp-content/uploads/sites/20/2019/03/JES-2018-4_47-56.pdf

Bradley, P. (2019). Risk management standards and the active management of malicious intent in artificial superintelligence. *AI & Society; London.* 35(2), 319-328. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2209789740?accountid=11774

Castelluccio, M. (2018a). The Malicious use of AI. *Strategic Finance; Montvale*. 99 (10), 55-57. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2062951395?accountid=11774

Castelluccio, M. (2018b). The Malicious use of AI: Part II. *Strategic Finance; Montvale*. 99 (11), 55-56. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2037357296?accountid=11774

Kilpatrick, H. (2019b). The Malicious Use of Artificial Intelligence in Cybersecurity. *Pipeline & Gas Journal; Dallas*. 246(2), 32. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2187375541/fulltextPDF/6418026A84D405DPQ/1?accountid=11774

Korolov, M. (2017). AI isn't just for the good guys anymore: Criminals are beginning to use artificial intelligence and machine learning to get around defenses. *CSO (Online); Framingham.* Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/1863568826?accountid=11774

Stephenson, P. (2018). Cloud-based security. *SC Magazine; New York*. 29(1), 29. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2002001269?accountid=11774

(f) Verkkosivu:

Beedham, M. (2020, 5.helmikuuta). Tesla's Autopilot dangerously fooled by drone-mounted projectors. Haettu 28.7.2020 osoitteesta https://thenextweb.com/cars/2020/02/05/teslas-autopilot-dangerously-fooled-by-drone-mounted-projectors/

Fralick, C. (2019, 15. tammikuu) Artificial Intelligence in Cybersecurity Is Vulnerable. Haettu 11.4.2020 osoitteesta https://www.scmagazine.com/home/opinion/artifical-intelligence-in-cybersecurity-is-vulnerable/

Halunen, K. (2018a, 19. helmikuuta). Lohkoketjusta tekoälyn luottamuksen rakentaja? Haettu 29.3.2020 osoitteesta https://vttblog.com/2018/02/19/lohkoketjusta-tekoalyn-luottamuksen-rakentaja/

Halunen, K. (2018b, 23. elokuuta). Tekoälykin voi haavoittua eikä täydellistä tekoälysovellusta ole. Haettu 12.1.2020 osoitteesta https://vttblog.com/2018/08/23/tekoalykin-voi-haavoittua-eika-taydellista-tekoalysovellusta-ole/

Halunen, K. (2019, 19. heinäkuuta). Miten tekoälyjä harhautetaan? Haettu 29.3.2020 osoitteesta https://www.vttresearch.com/fi/uutiset-ja-tarinat/miten-tekoalyja-harhautetaan

Independent. (2014, 1. toukokuuta). Haettu 8.7.2020 osoitteesta https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Kataja, M. (2020, 4. helmikuuta). Berliiniläistaiteilija loi virtuaalisen liikenneruuhkan – huijasi Google Mapsia kärryillä ja 99 kännykällä. Haettu 28.7.2020 osoitteesta https://yle.fi/uutiset/3-11191923

Kerns, J. (2017, 15. helmikuuta). What's the Difference Between Weak and Strong AI? Haettu 19.9.2020 osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/1876870051?accountid=11774

Kilpatrick, H. (2019a, 30. tammikuuta). How Artificial Intelligence Can Be Used Maliciously in Cybersecurity. Haettu 11.4.2020 osoitteesta https://www.datapine.com/blog/ai-in-cybersecurity/

Konttinen, M. (2018, 22. helmikuuta). Tutkijat maalaavat synkän kuvan tekoälyn vaaroista: Terroristit voivat kaapata robottiautoja ja miehittämättömiä ilma-aluksia iskujen tekemiseen. Haettu 31.3.2020 osoitteesta https://yle.fi/uutiset/3-10087215

Khurana, N., Mittal, S. & Joshi, A. (2018, 19. heinäkuuta). Preventing Poisoning Attacks on AI based Threat Intelligence Systems. arXiv.org; Ithaca. University of Maryland. Haettu osoitteesta https://search-proquest-com.ezproxy.jyu.fi/docview/2092804465?accountid=11774

Shen, J. & Xia, M. (2020, 30. heinäkuuta). AI Data poisoning attack: Manipulating game AI of Go. arXiv.org; Ithaca. Cornell University. Haettu osoitteesta https://arxiv.org/pdf/2007.11820.pdf

Lönnqvist, I. & Moilanen, P. (2017). Kyberin taskutieto. Keskeisin kybermaailmasta jokaiselle. Haettu 29.4.2020 osoitteesta https://jyx.jyu.fi/bitstream/handle/123456789/53510/978-951-39-7009-3.pdf?sequence=1&isAllowed=y

MC.AI. Aggregated news around AI and co. (2019, 4. helmikuu) 9 Critical AI Weaknesses to Consider. Haettu 5.4.2020 osoitteesta https://mc.ai/9-critical-ai-weaknesses-to-consider/

Mikrobitti. (2019, 2.huhtikuuta). Teslan autopilotin voi huijata vastaantulevien kaistalle tarran avulla – kuski ei huomaisi mitään. Haettu 28.7.2020 osoitteesta https://www.mikrobitti.fi/uutiset/teslan-autopilotin-voi-huijata-vastaantulevien-kaistalle-tarran-avulla-kuski-ei-huomaisi-mitaan/48925b26-3d3b-4b19-a5f6-20caa9f5e925 ja https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf

Mitchell, M. (2018, 1. kesäkuuta). U of T AI researchers design 'privacy filter' for photos that disables facial recognition systems. Haettu 5.4.2020 osoitteesta https://www.utoronto.ca/news/u-t-ai-researchers-design-privacy-filter-photos-disables-facial-recognition-systems

Roberts, S. (2018, 21. helmikuuta). Global AI experts sound the alarm. Leading researchers co-author unique report warning of the malicious use of AI in the coming decade. Haettu 19.9.2020 https://www.cam.ac.uk/Malicious-AI-Report

SAND Lab. (2020) Image "Cloaking" for Personal Privacy. Haettu 9.9.2020 osoitteesta http://sandlab.cs.uchicago.edu/fawkes/

STT. (2019, 9. syyskuuta). Tekoäly mullistaa vähitellen myös puolustusta ja sodankäyntiä – teknologian kehittyminen voi vähentää verenvuodatusta, mutta taustalla kummittelee pelko omin päin tappavista asejärjestelmistä. Haettu 31.3.2020 osoitteesta https://www.ksml.fi/kotimaa/Tekoäly-mullistaa-vähitellen-myös-puolustusta-ja-sodankäyntiä--%E2%80%89-teknologian-kehittyminen-voi-vähentää-verenvuodatusta-mutta-taustalla-kummittelee-pelko-omin-päin-tappavista-asej/1433385

Tarkoma, S. (2017). Tekoäly ja kokonaisturvallisuus. Maanpuolustus. Maanpuolustuskurssiyhdistyksen julkaisu. Haettu 30.3.2020 osoitteesta https://www.maanpuolustus-lehti.fi/single-post/Tekoaly-ja-kokonaisturvallisuus

Tivi. (2018, 1.kesäkuuta). Tekoälyt laitettiin taistelemaan – tuloksena uudenlainen suodatin kasvontunnistuksen hämäämiseen. Haettu 28.7.2020 osoitteesta https://www.tivi.fi/uutiset/tekoalyt-laitettiin-taistelemaan-tuloksena-uudenlainen-suodatin-kasvontunnistuksen-hamaamiseen/849c8525-ad00-3e07-9597-fee9295646d2

Townsend, K. (2018, 28. maaliskuu) The Malicious Use of Artificial Intelligence in Cybersecurity. Haettu 11.4.2020 osoitteesta https://www.securityweek.com/malicious-use-artificial-intelligence-cybersecurity

Zheng, C. (2017, 28. elokuu) The Cybersecurity Vulnerabilities to Artificial Intelligence. Haettu 5.4.2020 osoitteesta https://www.cfr.org/blog/cybersecurity-vulnerabilities-artificial-intelligence

(g) Raportti:

Ailisto, H. (toim.), Heikkilä, E., Helaakoski, H., Neuvonen, A. & Seppälä, T. (2018). *Tekoälyn kokonaiskuva ja osaamiskartoitus* (Julkaisusarjan osa 46). Valtioneuvoston kanslia.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. (2016). *Concrete Problems in AI Safety*. Cornell University.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R. & Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation and OpenAI.

Comiter, M. (2019). *Attacking Artificial Intelligence. AI's Security Vulnerability and What Policymakers Can Do About It*. Belfer Center for Science and International Affairs. Harvard Kennedy School.

European Comission. (2019). *Ethics guidelines for trustworthy AI.* High-Level Expert Group on Artificial Intelligence.

Floridi, L., Cowls, J., Beltrametti, M. Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, B. & Vayena, E. (2018). *AI4People's Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. The Scientific Committee of AI4People.

Patel, A., Hatzakis, T., Macnish, K., Ryan, M. & Kirichenko, A. (2019). *Security Issues, Dangers and Implications of Smart Information Systems*. Ref. Union's Horizon 2020 Research and Innovation Programme.

Tencent Keen Security Lab. (2019). Experimental Security Research of Tesla Autopilot.

Työ- ja elinkeinoministeriön julkaisuja. (2019). *Edelläkävijänä tekoälyaikaan*. Tekoälyohjelman loppuraportti. (Julkaisusarjan osa 23). Helsinki: Valtioneuvoston hallintoyksikkö.