Samuli Jylhä

SELITETTÄVÄT TEKOÄLYMALLIT (XAI) KYBERUHKIEN HAVAINNOINNISSA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2024

TIIVISTELMÄ

Jylhä, Samuli
Selitettävät tekoälymallit (XAI) kyberuhkien havainnoinnissa
Jyväskylä: Jyväskylän yliopisto, 2024, 117 s.
Kyberturvallisuus, pro gradu -tutkielma
Ohjaaja: Lehto, Martti

Tässä pro gradu -tutkielmassa tutkitaan selitettäviä tekoälymalleja kyberuhkien havainnoinnissa. Kybertoimintaympäristö on jatkuvassa murroksessa hyökkäysten monimutkaistuessa ja muuttuessa entistä vakavimmiksi. Hyökkäyksiä toteuttavien uhkatoimijoiden toimintaa pyritäänkin havaitsemaan mahdollisimman aikaisessa vaiheessa, jotta esimerkiksi arkaluonteisten tietojen varastaminen, kiristyshaittaohjelmien toteuttaminen tai muu vihamielinen toiminta kyettäisiin proaktiivisesti estämään. Kyberturvallisuuden alalla tekoälyä on otettu laajalti käyttöön erilaisten kone- ja syväoppimismallien muodossa. Suurin osa tekoälymalleista on kuitenkin niin monimutkaisia, ettei niiden toiminta ole läpinäkyvää tai selitettävissä.

Tutkimustehtävänä oli selvittää, miten selitettäviä tekoälymalleja voitaisiin hyödyntää kyberuhkien havainnoinnissa. Lisäksi tutkimuksessa pyrittiin löytämään vastauksia, voidaanko selitettävän tekoälyn avulla saavuttaa korkeampaa luotettavuutta ja tarkkuutta kyberturvallisuuden ratkaisuissa. Kolmantena tutkimusta tukevana lähestymiskulmana on, voisivatko kriittisien infrastruktuurin organisaatiot ja niihin vertautuvat instanssit, joilla on korkeat vaatimukset käytetyille tietojärjestelmille, hyötyä selitettävyydestä kyberturvallisuuden ratkaisuissa.

Pro gradu -tutkielman teoreettinen viitekehys koottiin edellä mainittujen tutkimuskysymysten näkökulmasta ja se taustoittaa lukijalle tutkittavaa aihetta ja siihen läheisimmin kytkeytyviä aihepiirejä. Tutkimus toteutettiin kvalitatiivisena tutkimuksena käyttäen aineistolähtöistä sisällönanalyysia. Tutkimusaineistoksi valikoitui kirjallisia dokumentteja sekä niiden tueksi yksi asiantuntijahaastattelu. Kirjallinen aineisto koostui alan tieteellisistä artikkeleista. Tutkimuksen tarkoituksena oli löytää ilmiöitä ja merkityksiä tutkimuskysymysten ympäriltä.

Sisällönanalyysin tuloksina tutkimusaineistoista muodostettiin kolme yhdistävää luokkaa, jotka olivat: ”XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa”, ”Perinteisten koneoppimismallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan” sekä XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla.”

Lopuksi muodostettiin johtopäätökset, joiden mukaan selitettävällä tekoäly parantaa tekoälymallin läpinäkyvyyttä ja sitä kautta luotettavuutta kyberturvallisuuden ratkaisuissa. Luotettavuus liittyy läheisesti tekoälymallien ja niihin kytkeytyvien sidosryhmien välille. Selitettävän tekoälyn tutkimukseen

kaivataan lisäksi vahvaa kontribuutiota sekä yrityksiltä, että tiedeyhteisöiltä. Havaittiin myös, että tutkimukseen on tärkeää liittää useita eri tieteenaloja.

Asiasanat: selitettävä tekoäly, XAI, kyberuhka, havainnointi, black-box, kyberturvallisuus

ABSTRACT

Jylhä, Samuli
Explainable artificial intelligence (XAI) models in cyber threat detection
Jyväskylä: University of Jyväskylä, 2024, 117 pp.
Cyber Security, Master's Thesis
Supervisor(s): Lehto, Martti

In this master´s thesis, explainable artificial intelligence models in cyber threat detection are researched. The cyber environment is in a constant state of flux as attacks become more complex and severe. The aim is to detect threat actors as early as possible in order to proactively prevent, for example, the theft of sensitive data, the execution of ransomware or other hostile actions. In the field of cyber security, AI has been widely deployed in the form of various machine and deep learning models. However, most AI models are so complex that their behaviour is not transparent or explainable.

The objective of this research was to investigate how AI models could be used for the detection of cyber threats. The research also aimed to find answers to the question of whether explainable AI can be used to achieve higher reliability and accuracy in cyber security solutions. A third approach that supports the research is whether critical infrastructure organisations and similar entities with high demands on the information systems they use could benefit from explainability in cybersecurity solutions.

The theoretical framework of the thesis was assembled from the perspective of the above-mentioned research tasks and provides the reader with a background to the topic under study and its most closely related themes. The research was conducted as a qualitative study using content analysis. The research material consisted of written documents and one interview with an expert of cyber security. The written material consists of scientific articles in the field. The aim of the study was to discover phenomena and meanings around the research questions.

As a result of the content analysis, three unifying categories were formed: "The demand for XAI models and the reinforcement they bring to AI solutions for cybersecurity", "The almost unreachable growing gap of traditional machine learning models in the race against new types of attacks" and "The growing need for research on XAI models in cybersecurity."

The study led to the conclusions that explainable AI improves the transparency of the AI model and thus its reliability in cybersecurity solutions. Reliability is closely related to the relationship between AI models and the stakeholders that are connected to them. However, humans are one of the most important elements in this equation. In addition, research on AI that can be explained requires a strong contribution from both the business and scientific communities. The importance of involving a wide range of disciplines in the research was also identified.

Keywords: explainable artificial intelligence, XAI, cyber threat, detection, black-box

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

1.1 Tutkimuksen taustaa

Kyberuhkat ovat jatkuvasti kasvussa oleva ongelma ja haaste eri organisaatioille sekä yrityksille. Tällaisiksi uhkiksi luetaan esimerkiksi kiristyshaittaohjelmat, valeuutiset ja kehittyneet haittaohjelmat, jotka voivat aiheuttaa merkittävääkin tuhoa tietojärjestelmissä. Organisaatiot kohtaavat kyberhyökkäyksiä enenevissä määrin, vaikka kyberturvallisuutta kehitetään jatkuvasti ja samaan aikaan aiheen ympärillä tehdään paljon tutkimusta. Lisäksi myös valtiot kehittävät kyberturvallisuuteen liittyviä kyvykkyyksiään. (Samtani ym., 2022.)

Kyberuhkien luonne on samanaikaisesti muuttunut vakavammaksi ja erilaisia hyökkäyksiä ja keinoja tunkeutua kohdejärjestelmiin tulee koko ajan enenevissä määrin lisää. On tärkeää pysyä hyökkääjien ja vihamielisten tahojen edellä puolustavan osapuolen näkökulmasta. Kyseessä onkin loputon kilpajuoksu, jossa kyberuhkien vaikutuksia pyritään lieventämään sekä havaitsemaan niitä ennen kuin ne ehtivät kunnolla realisoitua. (Samtani ym., 2022.)

Kyberhyökkäyksiä toteuttavien toimijoiden vaatimukset ovat olleet aiemmin huomattavasti korkeammat, ja hyökkäyksien valmistelemiseen ja tekemiseen on tarvittu paljon aikaa ja asiantuntijuutta. Samaan aikaan hyökkääjien käyttämät työkalut ovat olleet verrattain yksinkertaisia. Organisaatioiden kohtaamat uhkat ovat suuressa murroksessa. Kyberturvallisuuden toimintaympäristö on muuttunut merkittävästi viime vuosien aikana tekoälyn tullessa mukaan kentälle, jossa hyökkääjät ja puolustavat osapuolet yrittävät toteuttaa omia tehtäviään mahdollisimman laadukkaasti ja tehokkaasti. Hyökkääjät kykenevät tekoälyä hyödyntämällä automatisoimaan vaiheita, joita aiemmin on pitänyt toteuttaa käsin, mikä on samaan aikaan vaatinut paljon aikaa ja resursseja. Lisäksi tekoälyn avulla hyökkääjät kykenevät kehittämään työkaluistaan huomattavasti tehokkaampia, ja se mahdollistaa myös täysin uudenlaisia kyvykkyyksiä, joista puolustavien osapuolten ei ole tarvinnut aiemmin murehtia lainkaan. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022.)

Useat organisaatiot ovat implementoineet tekoälyratkaisuja kyberuhkien havainnointiin kasvavissa määrin viime aikoina. Sen avulla pyritään paranta-

maan jo olemassa olevia käytänteitä. Analysoitavan datan määrä on nykyään niin suuri, että tekoälyn on huomattu tuovan lisää tehoa ja kapasiteettia anomalioiden ja toistuvien kaavojen havaitsemisessa. (Samtani ym., 2022.)

Perustavanlaatuinen ongelma tekoälyn hyödyntämisessä on kuitenkin, että tekoälypohjaiset analytiikkatyökalut ovat niin kutsuttuja mustia laatikoita (engl. black-box), jotka eivät selitä käyttäjälleen, kuinka kyseinen malli päätyi tulokseen tai ennusteeseen, mihin se päätyi. Selittämättömyyden puutteella voi olla vakavia vaikutuksia ja seurauksia esimerkiksi tekoälyn luottamuksen suhteen. Tähän liittyvät huolet ovatkin aktivoineet tutkimusta aiheen ympärillä. On todettu, että on merkittävä tarve tutkia selitettäviä tekoälymalleja kyberturvallisuuden kentällä. (Samtani ym., 2022.)

1.2 Motivaatio

Tämän pro gradu -tutkielman yhtenä tärkeänä motivaatiotekijänä on, että datan määrä tulee todennäköisesti kasvamaan eksponentiaalisesti tulevaisuudessa ja lähitulevaisuudessa useilla aloilla. Tähän arvioon lukeutuu myös kyberturvallisuuden ala. Tekoälymallien hyödyntäminen suurten datamassojen analysoinnissa on siis välttämätöntä, mutta samanaikaisesti myös ongelmallista. Mallien implementointiin liittyy paljon avoimia kysymyksiä sekä suuria riskejä. Kuten edellisessä alaluvussa mainittua, niin myös kyberuhkien mittakaava ja vakavuus on kasvussa. Niihin vastaamiseen tarvitaan entistä järeämpiä ja luotettavampia ratkaisuja kyberturvallisuuden näkökulmasta.

Yhtenä analogiana voitaisiin tämän tutkielman tausta-ajatuksena käyttää seuraavaa: tavallisen kuluttajan selaillessa esimerkiksi verkkokauppaa, ja siellä tekoälyä hyödyntäen käyttäjälle suositellaan vahingossa tuotetta, joka ei herätä hänessä juurikaan kiinnostusta. Tästä ei todennäköisesti aiheudu esimerkiksi henkeä ja terveyttä vaarantavaa tai kansallista turvallisuutta horjuttavaa uhkaa.

Sen sijaan kybermaailmassa varsinkin kriittisen infrastruktuurin toimijoiden osalta tekoälyn tekemät virheet tai tekoälyn tuottamiin ratkaisuihin pohjautuvat päätökset voivat pahimmillaan aiheuttaa merkittäviä tuhoja monestakin eri näkökulmasta tarkasteltuna. Näin ollen tämän kaltaiset toimijat tarvitsevat todennäköisesti sellaisia tekoälymalleja kyberuhkien havainnointiin ja torjuntaan, jotka tarjoavat erityisen suurta luotettavuutta ja tarkkuutta. Selitettävillä tekoälymalleilla voisi olla mahdollisuuksia vastata tähän ongelmaan. Myös siihen tässä tutkielmassa pyritään löytämään vastauksia.

Tämän pro gradu -tutkielman yhtenä tärkeänä tavoitteena on aineistolähtöisen sisällönanalyysin avulla koota yhteen tutkimuskysymyksen alle osuvaa relevanttia aineistoa ja löytää ilmiöitä selitettävien tekoälymallien ja kyberturvallisuuden risteyskohdista. Tutkijan oma havainto on, että tekoälyn käyttö herättää paljon tunteita, voimakkaita mielipiteitä ja jopa pelkoa kyberturvallisuuden alalla. Näihin havaintoihin nivoutuvat myös indikaattorit selitettävyyden tarpeesta ja tekoälymallien kriittisen tarkastelun puutteesta.

1.3 Aikaisempi tutkimus

Tietokoneet ovat olleet osana ihmisten arkea vasta verrattain hyvin lyhyen aikaa ihmiskunnan historiassa. Kybermaailma käsitteenä ja siellä yksittäisiä ihmisiä sekä organisaatioita kohtaavat uhat ovat muuttuneet vakavimmiksi ja laaja-alaisempaa tuhoa aiheuttaviksi vasta viimeisen parin vuosikymmenen aikana. Näin ollen koko kyberturvallisuusympäristö ja siihen liittyvä tutkimus on uudehkoa.

Kyberturvallisuuden viitekehyksessä tekoälyn ja selitettävän tekoälyn hyödyntämisestä on alettu keskustella aktiivisemmin vasta muutamien viime vuosien aikana. Täten tieteellistä tutkimusta selitettävien tekoälymallien käytöstä kyberuhkien havainnoinnissa on tehty vielä vähän. Tutkimuksia on kuitenkin jo nähty, ja on havaittavissa, että kyberturvallisuusalalla kaupallisten yritysten, julkisen sektorin toimijoiden kuten myös akateemisen maailman sisällä tutkimusta tällä saralla tarvitaan runsaasti lisää.

Aiempaa tutkimusta, ja tämän tutkielman otsikon alle lukeutuvaa tutkimusta on koottu esimerkiksi Springerin julkaisemaan teokseen ”Explainable Artificial Intelligence for Cyber Security”, johon on koottu aihetta käsitteleviä artikkeleita (M. Ahmed ym., 2022).

Kyberuhkatiedusteluun liittyvää selitettävän tekoälyn tutkimusta on puolestaan koottu ”IEEE:n Transactions On Dependable And Secure Computing” -erikoisjulkaisuun (Samtani ym., 2022). Erikoisnumerossa esiteltyjä tieteellisiä artikkeleita on yhteensä seitsemän kappaletta. Kolme artikkelia käsittelee selitettäviä malleja sosiaalisen manipuloinnin ja tietojen kalastelun viitekehyksessä. Kaksi artikkelia keskittyy taas selitettäviin malleihin, joita hyödynnetään hyökkäysvektorien analysoimisessa. Kahdessa viimeisessä esitellyssä artikkelissa sen sijaan tutkitaan kyberpuolustuksen kehittämistä ja miten selitettäviä tekoälymalleja voitaisiin hyödyntää tässä kontekstissa. (Samtani ym., 2022.)

Lisäksi mainitsemisen arvoinen tutkimus on niin ikään IEEE:ssä on julkaistu kirjallisuuskatsaus vuodelta 2022, joka käsittelee nimenomaan selitettävää tekoälyä kyberturvallisuuden viitekehyksessä. Kyseisessä katsauksessa mainitaankin, että selitettävää tekoälyä ja kyberturvallisuutta on tutkittu, mutta lähinnä erikseen. Katsaus käsittelee selitettäviä tekoälytekniikoita yhdistettynä kyberturvallisuuden järjestelmiin. (Zhang, Hamadi, ym., 2022.)

1.4 Keskeiset käsitteet

Tässä luvussa tehdään läpileikkaus tämän pro gradu -tutkielman kannalta keskeisimpiin käsitteisiin. Tutkielmassa esiintyy lisäksi käsitteistöä, jota ei ole tässä avattu, mutta ne on pyritty selittämään erikseen lukijaa helpottamaan. Tutkimuksen aihealueen sisällä on paljon termistöä, jolle ei ole olemassa vakiintuneita suomenkielisiä termejä. Tutkielmassa on pyritty kääntämään kaikki termistö mahdollisimman osuvasti ja kuvaavasti. Tässä luvussa käsitteiden avaamisessa nojataan pääasiallisesti Suomen valtion turvallisuuskomitean laatimaan kyber-

turvallisuuden sanastoon (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018).

Riskillä tarkoitetaan epävarmuuden vaikutusta tavoitteisiin. Vaikutuksella tarkoitetaan tässä tapauksessa joko negatiivista tai positiivista vaikutusta. Samaan aikaan riski voi olla molempia. Riski voi kohdentua monien muiden kohteiden ohella tietojärjestelmiin. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 12.)

Yhteiskunnan elintärkeä toiminto viittaa sellaiseen toimintoon, jolla on välttämätön rooli yhteiskunnan toimivuuden näkökulmasta. Tähän lukeutuu esimerkiksi kansainvälinen- ja EU-toiminta, puolustuskyky, huoltovarmuus ja johtaminen. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 13.)

Kriittinen infrastruktuuri kuvaa niitä yhteiskunnan perusrakenteita, palveluita sekä niihin liittyviä toimintoja, jotka ovat välttämättömiä yhteiskunnan toimimisen ylläpidon kannalta. Sen alle nivoutuu sekä fyysisiä entiteettejä että digitaalisia toimintoja ja palveluita. Konkreettisia esimerkkejä ovat tieto- ja viestintäjärjestelmät sekä liikenne ja logistiikka. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 14.)

Tietoturva (engl. information security) tarkoittaa järjestelyitä, joiden avulla yritetään varmistaa, että tieto on saatavilla, eheää ja luottamuksellista. Saatavuudella viitataan siihen, että jokin tieto on hyödynnettävissä silloin kuin halutaan. Eheydellä tarkoitetaan, että tiedon sisältö pysyy samana verrattuna alkuperäiseen tietoon. Luottamuksellisuus on sitä, että tietoon pääsevät käsiksi vain valitut henkilöt. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 15.)

Haavoittuvuus (engl. vulnerability) on mikä vain heikkous, jonka avulla voidaan toteuttaa vahinkoa haluttuun kohteeseen. Tietojärjestelmät, prosessit ja ihmiset voivat sisältää haavoittuvuuksia. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 15.)

Kyber (engl. cyber) -sanaa hyödynnetään usein alan yhdyssanoissa. Kyber -sana liitetään useimmissa käyttötapauksissa sellaisen informaation käsittelyyn, joka on digitaalisessa muodossa. Tällaiset yhdyssanat kytkeytyvät tietotekniikkaan, digitaaliseen viestintään, tietoverkkoihin, tietojärjestelmiin ja tietokonejärjestelmiin. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 21.)

Kybertoimintaympäristö (engl. cyber domain) on sellainen ympäristö, joka koostuu vähintään yhdestä digitaalisesta tietojärjestelmästä. Järjestelmiä voi olla, ja usein onkin useampia kuin yksi. Kybertoimintaympäristö hyödyntää elektroniikkaa sekä sähkömagneettista spektriä datan ja informaation varastoimiseen, muokkaukseen sekä siirtoon. Kaikki tämä tapahtuu viestintäverkkoja käyttäen. Myös sellaiset fyysiset rakenteet ovat osa kybertoimintaympäristöä, jotka liittyvät datan ja informaation käsittelyyn. Liikenteen ohjausjärjestelmät, tietojärjestelmiin pohjautuvat ydinvoimalan ohjausjärjestelmät sekä pankki- ja maksujärjestelmät ovat havainnollistavia esimerkkejä erilaisista kybertoimintaympäristöistä. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 21.)

Kyberturvallisuus (engl. cyber security) on käsite, joka kuvastaa tavoitetilaa, jossa luottamus on syntynyt kybertoimintaympäristöä kohtaan. Samaan aikaan se tarkoittaa kybertoimintaympäristön turvaamista toimenpiteillä, joilla voidaan proaktiivisesti hallita kyberuhkia. Yhtäältä kyberuhkia ja niiden vaikutuksia pitää myös sietää. Suomen kybertoimintaympäristön haasteisiin vastaamista

ja sen toimivuuden varmistamista ohjaa Suomen kyberturvallisuusstrategia. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 22.)

Kyberpuolustus (engl. cyber defense) on osa-alue, joka kytkeytyy maanpuolustukseen. Se pitää sisällään tiedustelun, vaikuttamisen sekä suojautumisen suorituskyvyt. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 22)

Tietoturvauhka (engl. data security threat) on haitallinen tapahtuma tai kehityskulku, joka toteutuu mahdollisesti. Sen kohteena on tietoturva. Uhkan realisoituessa, tietoturva vaarantuu. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 25.)

Kyberuhka (engl. cyber threat) on puolestaan käsite, jolla tarkoitetaan samoin kuin tieturvauhkan tapauksessa, mahdollisesti realisoituvaa tapahtumaa tai kehityskulkua, jonka kohteena on kybertoimintaympäristö. Jos kyberuhka toteutuu, vaarantaa se toiminnot, jotka ovat siitä riippuvaisia. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 25.)

Uhkatoimija (engl. threat actor) on taho tai toimija, joka on vastuussa tapahtumasta, joka vaikuttaa tai mahdollisesti vaikuttaa jonkin organisaation turvallisuuteen (Sailio ym., 2020, s. 2). Termistä käytetään myös vaihtoehtoisesti käsitettä *kyberuhkatoimija (engl. cyber threat actor).* Uhkatoimijoiden koko vaihtelee yksittäisten henkilöiden ja isompien ryhmien välillä. Niille yhteistä on tahallisen vahingon aiheuttaminen digitaalisille järjestelmille ja laitteille. Uhkatoimijat toimivat hyväksikäyttäen tietojärjestelmien, verkkojen sekä ohjelmistojen haavoittuvuuksia. Uhkatoimijoiden kirjo on tänä päivänä laaja ja kaikilla toimijoilla on erityyppiset taktiikat, tekniikat, menetelmät, taitotasot ja motiivit. Esimerkkejä uhkatoimijoista ovat hakkeriaktivistit (haktivistit), valtiolliset toimijat, kyberrikolliset, sisäpiiriläiset sekä kyberterroristit. (*What Is a Threat Actor?*, 2024.)

Kyberrikollisuus (engl. cybercrime) on käsite, jolla tarkoitetaan sellaista rikollisuutta, jossa hyödynnetään viestintäverkkoja sekä tietojärjestelmiä. Se voi myös kohdistua näihin. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 26.)

Attribuutio (engl. attribution) on termi, jolla viitataan kyberhyökkäyksen tehneen toimijan tunnistamiseen. Lisäksi tähän voi sisältyä toimijan paikantaminen ja mahdollisesti oikeudelliseen vastuuseen saattaminen. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 27)

Kyberhyökkäys (engl. cyber attack) on teko tai toiminta, joka tapahtuu tietoverkon kautta. Sen avulla pyritään datan vahingoittamiseen tai käyttöön, johon hyökkäyksen toteuttajalla ei ole oikeutta. Kyberhyökkäyksen toimet voivat kohdistua myös tietoverkkoihin -ja järjestelmiin, kuten myös fyysisiin laitteisiin. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 30)

Palvelunestohyökkäys (engl. denial of service) on tietoverkon hyökkäys, jonka tarkoituksena on aiheuttaa kohdeverkkoon- tai järjestelmään niin suuri verkkoliikenteen ruuhka, että se saa jonkin halutun palvelun lamaantumaan. Kohteena voi olla esimerkiksi jokin verkkosivusto. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 31.)

APT-hyökkäys (engl. advanced persistent threat, APT-attack) on käsite, jolla viitataan kohdistettuun kyberhyökkäykseen, ja joka kohdistuu tiettyyn ennalta valikoituun kohteeseen. Sen toteuttamiseen käytetään haittaohjelmia ja muita vastaavia toimintoja. Tällaiset hyökkäykset ovat monivaiheisia, ja niissä pyri-

tään pitkäkestoisuuteen. Haittaohjelmat, joita niissä käytetään, voivat olla yksilöllisesti räätälöityjä. APT-ryhmillä tarkoitetaan itsenäisesti tai vaihtoehtoisesti valtion alla toimivaa organisoitunutta hakkeriryhmää. APT-hyökkäykset voivat kohdistua esimerkiksi yrityksiin, valtionhallintoihin tai niiden alla toimiviin organisaatioihin sekä valittuihin henkilöihin. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 31.)

Haittaohjelma (engl. malware) on sellainen tietokoneohjelma, joka aiheuttaa tahallisesti ei-toivottuja toimia tietojärjestelmässä. Tällaiset tapahtumat voivat olla haitallisia laitteen käyttäjää tai tietojärjestelmää kohtaan. Ne voivat olla myös molempia. (*Kyberturvallisuuden sanasto – Turvallisuuskomitea*, 2018, s. 31.)

Kyberuhkatiedustelulla (engl. cyber threat intelligence, CTI) tarkoitetaan prosessia sekä tuotetta, joka jalostuu datasta analyysin kautta tiedoksi. Tämän kyberuhkatiedustelun prosessin tuotoksena syntyvän tuotteen tulee täyttää vaatimukset, jotka liittyvät vastustajiin ja uhkatoimijoihin, joilla on halu, kyky sekä mahdollisuus toteuttaa vahingollisia toimia kybertoimintaympäristössä. Toimialana kyberuhkatiedustelu on johdettu vaaratilanteisiin vastaamisesta sekä perinteisestä tiedustelusta. Kyberuhkatiedustelu on viimeisen kymmenen vuoden aikana muotoutunut ja löytänyt paikkansa kriittisenä osana organisaatioiden tietoturvaan liittyvien operaatioiden valmiuksia. (*Exploring the Opportunities and Limitations of Current Threat Intelligence Platforms*, ei pvm., s. 7.)

Datan keräys kohdistuu tietoturvauhkiin, uhkatoimijoihin, hyväksikäyttöihin, haittaohjelmiin, haavoittuvuuksiin ja vaarantuneisiin indikaattoreihin (ip-osoitteet, domainit ja niin edelleen). Kerätystä datasta tehdään arvioita ja sitä sovelletaan organisaation tarpeiden mukaan. Kyberuhkatiedustelu on ajan saatossa muodostunut kyberturvallisuuden alalla toimivien asiantuntijoiden työkaluksi, jonka avulla vastataan hyökkäyksiin tarkasti ja oikea-aikaisesti. (Conti ym., 2018, s. 2.)

Tekoäly (engl. artificial ontelligence, AI) on käyttäytymisen ymmärtämiseen liittyvä lähestymistapa, joka pohjautuu olettamaan, että älykkyyttä voitaisiin analysoida parhaiten sen toistamisella. Tässä kontekstissa toistamisella viitataan tietokoneella tehtävään simulointiin. Historian valossa tekoäly on suhteellisen tuore käsite ja se on muodostunut itsenäisenä tutkimusalana alun perin 1950-luvun puolivälissä. Tekoälyn katsotaan olevan osa tietojenkäsittelytieteitä. (Garnham, 2018, s. 2.) Tekoälyä pidetään tänä päivänä niin kutsutun neljännen sukupolven teollisuuteen (engl. Industry 4.0) siirtymisen suurimpana muutosvoimana. Tekoälyn avulla älykkäät koneet kykenevät suorittamaan erilaisia tehtäviä täysin itsenäisesti, mihin lukeutuu esimerkiksi datan analysointi (I. Ahmed, 2022).

Selitettävällä tekoälyllä (engl. explainable artificial intelligence, XAI) tarkoitetaan tekoälymalleja, jotka ovat algoritmien ja työkalujensa puolesta malleja, joiden prosesseja ja päätöksiä ihminen pystyy ymmärtämään. Ne ovat siis toinen ääripää aiemmin kuvailluille black-box -malleille. (Ahmed, 2022.)

Tässä pro gradu tutkielmassa pyritään pääasiallisesti käyttämään termiä selitettävä tai selittävä tekoäly. Joissain yhteyksissä tullaan kuitenkin käyttämään pelkkää ”*XAI*” -lyhennettä.

1.5 Tutkimuksen keskeiset tulokset

Tutkimuksen keskeisinä tuloksina sisällönanalyysin pohjalta muodostettiin kolme yhdistävää luokkaa tutkimusaineistosta. Yhdistävä luokat ovat:

- XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa,
- perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan,
- XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla.

1.6 Tutkimuksen rakenne

Tämä tutkimus koostuu kuudesta eri pääluvusta. Johdantoluvussa lukija johdatellaan aihealueen sisään. Johdannossa on kuvattu tutkimuksen taustat, motivaatio, aiempi tutkimus, keskeiset käsitteet sekä tutkimuksen keskeisimmät tulokset.

Toisessa luvussa syvennytään tutkimuksen toteutukseen. Siinä tehdään kattava kuvaus tutkimuksen lähtökohtiin ja tutkimuskysymyksiin. Sen jälkeen siirrytään tarkastelemaan, kuinka tutkimus on tarkennettu ja rajattu. Luvun päätteeksi avataan tutkimuksessa käytetyt tutkimusmenetelmät sekä aineistonkeruutavat. Käytetty tutkimusmenetelmä kuvataan tarkasti läpi vaihe vaiheelta konkreettisin ja havainnollistavin esimerkein.

Kolmannessa ja neljännessä luvuissa luodaan tutkimuksen kannalta merkityksellinen viitekehys, jotta lukija saa peruskäsityksen tutkittavasta aiheesta ja ilmiöistä. Lähdeaineisto, johon viitekehys pohjautuu, on pyritty kokoamaan käyttämällä mahdollisimman laajasti erilaista alan kansainvälistä kirjallisuutta. Lähdeaineisto painottuu vertaisarvioituihin tieteellisiin artikkeleihin sekä kirjoihin. Lähteinä on käytetty lisäksi raportteja sekä web-sivustoja. Lähteiden hankintaan pääasiallisena tietokantana käytettiin Jyväskylän yliopiston kirjaston JYKDOK -tietokantaa sen kattavuuden ja luotettavuuden vuoksi. Sen lisäksi käytettiin eri hakukoneita, kuten Googlea ja Bingiä. Lisäksi hyödynnettiin Google Scholaria.

Viidennessä luvussa käsitellään tutkimuksen tulokset. Kuudennessa luvussa esitetään pohdinta peilaten tutkimuksen tuloksia sekä teoreettista viitekehystä toisiinsa. Lopuksi avataan tutkimuksen johtopäätökset.

2 TUTKIMUKSEN TOTEUTUS

2.1 Tutkimuksen lähtökohta ja tutkimuskysymykset

Tutkimuksen tavoitteena on selvittää, kuinka selitettäviä tekoälymalleja voisi hyödyntää kyberuhkien havaitsemisessa eri kyberturvallisuuden ratkaisuissa. Lisäksi on määritelty tarvittavat alatutkimuskysymykset. Taustalla on ajatus, että mikäli selitettävien mallien avulla päästäisiin korkeampaan luotettavuuteen, laskisiko se korkean varautumisen organisaatioiden kynnystä implementoida tällaisia malleja heidän käyttöönsä. Tähän segmenttiin luetaan esimerkiksi kriittisen infrastruktuurin kannalta merkitykselliset toimijat.

Tutkimuksen tavoitteen perustella tutkimukselle määritettiin yksi päätutkimuskysymys, joka on listan ensimmäisenä. Tueksi muodostettiin kaksi alatutkimuskysymystä, jotka ovat jäljempänä:

- Miten selitettäviä tekoälymalleja voisi hyödyntää kyberuhkien havaitsemisessa?
- Voidaanko selitettävillä tekoälymalleilla saavuttaa korkeampaa luotettavuutta kyberturvallisuuden ratkaisuissa?
- Kuinka selitettävät mallit soveltuisivat organisaatioille, jotka vaativat korkeaa laatua ja tarkkuutta käytettäviltä tietojärjestelmiltä?

2.2 Tarkennus ja rajaus

Tutkimuksen aihetta on tutkittu jonkin verran, mutta viitekehys on vielä verrattain uusi. Tavoitteena on rajata selitettävien tekoälymallien hyödyntämistä kyberhyökkäyksiä kuvaavien hyökkäysketjumallein ensimmäisiin vaiheisiin. Kyberhyökkäykset havaitaan usein vasta silloin, kun ne ovat jo tapahtuneet ja vahinkoa on päässyt käymään.

Rajaus on kuitenkin tästä näkökulmasta haasteellista, koska lähde- ja tutkimusaineistoa aiheesta on vielä yleensäkin vähän. Spesifejä tutkimuksia hyök-

käysketjun ensimmäisiin vaiheisiin liittyen tutkimuksen kontekstissa ei ole juurikaan löydettävissä. Täten tutkimuksen kärjen kaventaminen vain hyökkäysketjumallien ensimmäisiin vaiheisiin voi muodostua vaikeaksi, mutta siitä huolimatta tässä tutkielmassa yritetään löytää vastauksia tästä näkökulmasta.

Uhkatoimijoiden toteuttamia tiedustelutoimenpiteitä organisaatioita ja niiden tietojärjestelmiä kohtaan yritetäänkin kyberturvallisuusalan ammatinharjoittajien ja asiantuntijoiden toimesta havaita mahdollisimman aikaisessa vaiheessa ennen kuin mitään pahaa niin sanotusti ehtii tapahtua. Tämä on luonnollisesti sitä haastavampaa, mitä enemmän liikutaan hyökkäysketjujen ensimmäisiin vaiheisiin. Tavoite on löytää vastauksia tähän ongelmaan ja haasteeseen tämän tutkimuksen avulla.

2.3 Tutkimusmenetelmät ja aineistonkeruutavat

Tutkimus toteutetaan kvalitatiivisena tutkimuksena käyttäen aineistolähtöistä sisällönanalyysia. Kvalitatiivisessa tutkimuksessa yksi perusperiaate on kuvata todellista elämää ja ilmiöitä. Samaan aikaan pyritään, että tutkittavaa kohdetta tutkittaisiin mahdollisimman kokonaisvaltaisesti. On yleisesti todettua, että kvalitatiivisen tutkimuksen yksi tavoite on löytää tai nostaa esiin tosiasioita sen sijaan, että yritettäisiin osoittaa todeksi jo tiedettyjä tai olemassa olevia totuuksia. (Hirsjärvi ym., 2009, s. 161.)

Kvalitatiivisella eli laadullisella tutkimuksella on useita piirteitä, joilla tutkimusmenetelmää voidaan kuvata. Niihin lukeutuvat kokonaisvaltainen tutkimuksen tekeminen ja tiedon hankkiminen, jossa tutkittava aineisto pyritään kokoamaan luonnollisissa ja todellisissa tilanteissa. Samalla yhtenä kvalitatiivisen tutkimuksen tunnusmerkkinä pidetään ihmistä tiedon keräämisen instrumenttina. (Hirsjärvi ym., 2009, s. 164.)

Tämän tutkielman kvalitatiivisen tutkimuksen lajina käytetään aineistolähtöistä sisällönanalyysia, jolloin kerätty aineisto koostuu pääasiassa kirjallisesta aineistosta. Ihminen on luonnollisesti aina myös kirjalliseen muotoon saatetun tiedon takana.

Hirsjärvi ym. (2009, s. 164) listaavat kvalitatiivisen tutkimuksen tunnusmerkistöön myös induktiivisten analyysien käyttämisen, jolla viitataan tutkijan pyrkimyksiin paljastaa odottamattomia seikkoja. Tällöin ei lähdetä testaamaan eri teorioita tai hypoteeseja, vaan sen sijaan aineistoa yritetään tarkastella monitahoisesti ja yksityiskohtaisesti. Kvalitatiivisessa tutkimuksessa suositaan myös laadullisten metodien käyttöä aineiston hankinnassa eli tavoitellaan, että tutkittavien näkökulmat pääsisivät mahdollisimman hyvin esille. Tämän tyyppisiin metodeihin lukeutuvat esimerkiksi haastattelut sekä dokumenttien ja tekstien diskursiivinen analysointi. Tutkimussuunnitelma ei ole myöskään niin sanotusti lopullinen laadullista tutkimusta tehtäessä. Se voi muotoutua tutkimusta tehdessä ja sen edetessä. Suunnitelmia voidaan siis tarvittaessa muuttaa olosuhteiden ja tarpeiden mukaisesti. (Hirsjärvi ym., 2009, s. 164.)

Kun tehdään laadullista tutkimusta, yleisimpiin aineiston keräysmenetelmiin sisältyvät esimerkiksi haastattelut, kyselyt sekä erilaisista dokumenteista

kerätty tieto. Erityyppisiä menetelmiä voidaan hyödyntää joko rinnakkain tai eri tyyleillä yhdistelemällä riippuen siitä, mikä tutkittava ongelma on ja kuinka paljon tutkimukseen on käytettävissä resursseja. (Tuomi & Sarajärvi, 2018, s. 83.)

Kerätyn aineiston analysoimiseen käytetään sisällönanalyysia ja tarkemmin aineistolähtöistä sisällönanalyysia. Sisällönanalyysille tyypillistä ja tunnusomaista on, että sen avulla voidaan analysoida dokumentteja systemaattisesti ja objektiivisesti. Dokumenttia ei ole määritelty kyseisessä asiayhteydessä kovinkaan tiukasti. Sellaisiksi voidaan nähdä esimerkiksi artikkelit, kirjat, haastattelut, puheet, raportit ja niin edelleen. Miltei mikä tahansa materiaali, joka on kirjallisessa muodossa, voidaan katsoa olevan dokumentti. Sisällönanalyysin yhtenä etuna on mainittu myös, että se soveltuu myös sellaisen aineiston analysointiin, jota ei ole strukturoitu millään tavalla. Sisällönanalyysin avulla tavoitteena on tuottaa kuvaus tutkittavasta ilmiöstä tiivistetyssä sekä yleisessä muodossa. (Tuomi & Sarajärvi, 2018, s. 117.)

Tässä pro gradu -tutkielmassa nojaudutaan pääasiassa aineiston keräämisen osalta kirjallisten dokumenttien keräämiseen. Kirjallisen aineiston tukena tullaan käyttämään yhtä asiantuntijahaastattelua. Dokumentteja kerätään mahdollisimman laajasti erilaisista lähteistä. Tutkittavasta aiheesta on tehty melko vähän tieteellistä tutkimusta, mutta tieteellisiä artikkeleita on muutaman viime vuoden aikana tuotettu tiedeyhteisössä jo jonkin verran.

Haastattelu toteutettiin strukturoimattomana syvähaastatteluna. Tämän tyyppisestä haastattelusta käytetään myös nimitystä avoin haastattelu tai keskustelunomainen haastattelu. Tämän kaltaisessa haastattelussa käytetään pääasiassa avoimia kysymyksiä. Ainoa etukäteen määritelty asia on ilmiö, josta keskustellaan. Syvähaastattelu ei kuitenkaan tarkoita sitä, että vain avoimia kysymyksiä esittämällä siitä muodostuisi sellainen. Sen sijaan haastattelijalla on tässä hyvin suuri vastuu ja tehtävä syventää haastattelua sen aikana saatujen vastausten perusteella. Syvähaastattelussa tunnusomaista on tutkittavan ilmiön perinpohjainen avaaminen, jolloin haastateltavia voi olla vain yksi. On myös mahdollista, että samaa henkilö voidaan haastatella useammin kuin kerran. (Tuomi & Sarajärvi, 2018, s. 88).

Haastattelu olisi voitu toteuttaa myös lomakehaastatteluna tai puolistrukturoituna teemahaastatteluna, mutta tämän tutkimusongelman ratkaisemiseksi soveltuu parhaiten juuri syvähaastattelu. Teemahaastattelun ja syvähaastattelun asetelmissa on perustavanlaatuinen ero. Teemahaastattelussa oletetaan haastateltavien ymmärtävän ja tulkitsevan jonkin tutkimuksen teoreettisessa osassa käsitellyn asian tismalleen tavalla, jolla se on esitetty. Teemahaastatteluissa oletetaan lisäksi, että haastateltavat ovat kyvykkäitä jäsentelemään tällaisen asian samalla tavalla kuin tutkija. (Tuomi & Sarajärvi, 2018, s. 89–90.)

Tämän pro gradu -tutkielman aihe on sen verran tuntematon, joten teemahaastattelu voisi olla tästäkin näkökulmasta haastavaa toteuttaa ja se voisi kaatua tässä tapauksessa omaan kankeuteensa. Avoimen haastattelun avulla haastattelua voidaan tutkijan toimesta modifioida sen edetessä tarpeen mukaan ja sitä kautta päästä syvällisempään lopputulokseen aineiston analysoinnin ja johtopäätösten tekemisessä.

2.3.1 Kirjallisen aineiston keräys

Sisällönanalyysia varten tähän tutkielmaan pyrittiin löytämään mahdollisimman osuvaa ja relevanttia aineistoa, joka käsittelee selitettäviä tekoälymalleja juuri kyberturvallisuuden viitekehyksessä. Aineistoa pyrittiin löytämään myös kyberuhkatiedustelun (engl. cyber threat intelligence, CTI) alueelta. Aineiston haku kohdennettiin seuraaviin tietokantoihin:

- Jyväskylän yliopiston kirjaston JYKDOK,
- IEEE,
- DOAJ,
- Scopus,
- Google Scholar.

Hakusanoina käytettiin seuraavia englannin kielen sanoja: explainable, artificial, intelligence ja cyber. Haut toteutettiin käyttämällä hakuoperaattoria ”AND” haettavien sanojen välissä: explainable AND artificial AND intelligence AND cyber. Haku kohdistettiin otsikoihin, jotta saataisiin rajattua tuloksia mahdollisimman hyvin.

Tutkimukseen valikoitiin kirjalliseksi aineistoksi tieteellisiä artikkeleita. Keräykseen tehtiin rajaus julkaisuajankohdan mukaan. Aineistoon ei valikoitu ennen vuotta 2021 julkaistuja artikkeleita. Rajaus tehtiin sen vuoksi, koska tässä tutkielmassa käsitelty aihe on tieteen kentällä hyvin uusi ja relevanttia aineistoa olisi ollut todennäköisesti haastavaa löytää 2010-luvulta. Lisäksi artikkelit valikoitiin siten, että ne olisivat vertaisarvioituja. Haun tuottamat tulokset tietokannoittain:

- JYKDOK: 28 kpl,
- IEEE: 3 kpl,
- DOAJ: 2 kpl,
- Scopus: 8 kpl,
- Google Scholar: 27 kpl.

On tärkeää huomioida, että hauissa tuli päällekkäisiä tuloksia jonkin verran. Samalla voidaan tehdä havainto, että tieteellisiä artikkeleita selitettävästä tekoälystä kyberturvallisuuden kentällä on vielä vähän. Tämä helpotti toisaalta tutkijaa siinä mielessä, että aineiston valikointi oli helpompaa kuin jos hakutuloksia olisi ollut esimerkiksi useita satoja. Kirjalliseksi aineistoksi valikoidut artikkelit on esitetty taulukossa 1.

TAULUKKO 1 Analysoitava kirjallinen aineisto.

Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Phishing Email Detection Using Persuasion Cues Phishing-sähköpostin havaitseminen suostutteluvihjeiden avulla	Rohit Valecha , Pranali Mandaokar, and H. Raghav Rao	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
XAI-Based Microarchitectural Side-Channel Analysis for Website Fingerprinting Attacks and Defenses XAI-pohjainen mikroarkkitehtuurin sivukanava-analyysi verkkosivujen sormenjälkihyökkäyksiä ja puolustusta varten	Berk Gulmezoglu	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Explainable Intelligence-Driven Defense Mechanism Against Advanced Persistent Threats: A Joint Edge Game and AI Approach Selitettävissä oleva tiedusteluun perustuva puolustusmekanismi kehittyneitä pysyviä uhkia vastaan: Edge-pelin ja tekoälyn yhteinen lähestymistapa	Huiling Li , Jun Wu , Member, IEEE, Hansong Xu , Gaolei Li , and Mohsen Guizani , Fellow, IEEE	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence Selitettävissä oleva monimodaalinen hierarkkinen tarkkaavaisuusmalli phishing-uhkatiedustelun kehittämiseksi.	Yidong Chai, Yonghang Zhou, Weifeng Li, and Yuanchun Jiang	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Open Source Intelligence for Malicious Behavior Discovery and Interpretation Avointen lähtieden tiedustelu haitallisten toimintatapojen havaitsemiseen ja tulkintaan	Yi-Ting Huang , Chi Yu Lin , Ying-Ren Guo , Kai-Chieh Lo, Yeali S. Sun and Meng Chang Chen	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
The Past, Present, and Prospective Future of XAI: A Comprehensive Review XAI:n menneisyys, nykyisyys ja tulevaisuudennäkymät: kattava katsaus	Muhammad Usama Islam, Md. Mozaharul Mottalib, Mehedi Hassan, Zubair Ibne Alam, S. M. Zobaed & Md. Fazle Rabby	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform Selitettävä tekoäly älykkäiden kaupunkien sovelluksia varten: Turvallinen ja luotettava alusta	M. Humayun Kabir, Khondokar Fida Hasan, Mohammad Kamrul Hasan, and Keyvan Ansari	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Explainable Artificial Intelligence in Sustainable Smart Healthcare Selitettävä tekoäly kestävässä älykkäässä terveydenhuollossa	Mohiuddin Ahmed and Shahrin Zubair	Tieteellinen artikkeli	2022
Aineisto/dokumentti	Tekijät	Tyyppi	Vuosi
Adversarial XAI Methods in Cybersecurity Adversariaaliset XAI-menetelmät kyberturvallisuudessa	Aditya Kuppa and Nhien-An Le-Khac	Tieteellinen artikkeli	2021

Tässä lisäksi tutkimusaineisto luetelman muodossa:

- Phishing Email Detection Using Persuasion Cues (Valecha ym., 2022),
- XAI-Based Microarchitectural Side-Channel Analysis for Website Fingerprinting Attacks and Defenses (Gulmezoglu, 2022)
- Explainable Intelligence-Driven Defense Mechanism Against Advanced Persistent Threats: A Joint Edge Game and AI Approach (Li ym., 2022),
- An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence (Chai ym., 2022),
- Open Source Intelligence for Malicious Behavior Discovery and Interpretation (Huang ym., 2022),
- The Past, Present, and Prospective Future of XAI: A Comprehensive Review (Islam ym., 2022),
- Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform (Kabir ym., 2022),
- Explainable Artificial Intelligence in Sustainable Smart Healthcare (M. Ahmed & Zubair, 2022),
- Adversarial XAI Methods in Cybersecurity (Kuppa & Le-Khac, 2021)
- Asiantuntijahaastattelu (*Asiantuntijahaastattelu*, henkilökohtainen viestintä, 16. helmikuuta 2024).

2.3.2 Asiantuntijahaastattelu

Kuten aiemmin mainittua, tutkimuksen kirjallisen aineiston tueksi tehtiin yksi asiantuntijahaastattelu, joka toteutettiin strukturoimattomana haastatteluna. Haastatteluun valittiin kaupallisesta suomalaisesta tietoturvayhtiöstä asiantuntija, jolla on laaja-alaista kokemusta ja näkemystä kyberturvallisuuden eri osa-

alueilta. Haastatteluun valitulla asiantuntijalla ei ollut kokemusta selitettävistä tekoälymalleista entuudestaan, mikä oli yksi syy sille, että haastattelu toteutettiin rakenteeltaan avoimena ja mahdollisimman strukturoimattomana. Haastattelulle olisi ollut ongelmallista rakentaa struktuuria, jos ei tiedetä, kuinka paljon haastateltava tietää aiheesta.

Tutkijan havaintojen mukaan selitettävä tekoäly on kyberturvallisuuden kentällä vielä niin uusi ja tuntematon aihealue, että asiantuntijaa, jolla olisi siitä kattavaa kokemusta, olisi ollut hyvin hankala löytää tähän pro gradu -tutkielmaan haastateltavaksi.

Haastateltavaa lähestyttiin sähköpostitse ja haastattelu sovittiin pidettäväksi asiantuntijan edustaman yrityksen toimitiloissa. Haastattelu toteutettiin siis kasvotusten ja sille varattiin aikaa puolitoista tuntia. Aika käytettiin täysimääräisesti. Haastattelu toteutettiin helmikuussa 2024.

Avoin haastattelu toimi tässä tapauksessa hyvin, koska haastattelua pystyi mukauttamaan sen edetessä. Samalla avoimen haastattelun toteuttaminen oli myös haastavaa, koska haastateltavan vastauksien perusteella piti kyetä etenemään nopeasti ja mahdollisimman luontevasti aina seuraavaan kysymykseen. Haastattelua varten oli kuitenkin laadittu muutamia tukikysymyksiä etukäteen.

Haastattelun alkuun oli myös mietitty valmiiksi muutamia valmistelevia kysymyksiä, joiden tarkoituksena oli saada keskustelu sulavasti käyntiin ja samalla rakentaa luottamusta haasteltavan ja haastattelijan välillä. Vaikka haastateltavalla ei ollut suoranaista käytännön kokemusta selitettävistä tekoälymalleista, niin haastattelussa saatiin siitä huolimatta toivotunlaista keskustelua ja näkemyksiä aiheesta sekä sen ympäriltä. Aineiston keruu onnistui haastattelun osalta siis ennalta suunnitellusti.

2.3.3 Sisällönanalyysi

Tähän pro gradu -tutkielmaan kerätty tutkimusaineisto analysoitiin käyttämällä aineistolähtöistä sisällönanalyysia. Seuraavassa kuvataan analyysiprosessin vaiheet ja kuinka tätä työtä varten kerätty aineisto analysoitiin kyseistä tutkimusmenetelmää käyttäen.

Aineistolähtöisestä laadullisesta analyysista käytetään myös ilmausta induktiivinen analyysi. Tässä työssä puhutaan kuitenkin aineistolähtöisestä analyysista. Sen prosessi on kolmivaiheinen:

- Aineiston redusointi (pelkistäminen),
- Aineiston klusterointi (ryhmittely),
- Abstrahointi (teoreettisten käsitteiden luominen) (Tuomi & Sarajärvi, 2018, s. 122.)

Sisällönanalyysin aloittamista edeltää analyysiyksikön määrittäminen. Analyysiyksiköksi voidaan määrittää yksittäinen sana, lause, lausuma tai ajatuskokonaisuus. Ajatuskokonaisuuden sisään voi asettua useampia lauseita. Kun analyysiyksikköä lähdetään määrittelemään, täytyy tutkimustehtävän ja aineiston laadun ohjata määrittelyä. (Tuomi & Sarajärvi, 2018, s. 122.)

Analyysiyksiköksi valittiin yksittäiset sanat, jotka kuvaavat selitettävyyttä, tulkittavuutta ja kyberuhkien havainnointia sekä niihin vastaamista. Myös tekoälyyn ja sen alle nivoutuvat keskeiset termit valikoitiin analyysiyksiköiksi. Kaikki kirjallinen aineisto oli englanninkielistä, joten analyysiyksiköiksi valikoitu niin ikään englanninkielen termit: *explainable, interpretable, artificial intelligence, machine learning, deep learning, observation, detection, trust, accuracy, decision, cyber threat, cyber threat intelligence, kill chain ja black-box.*

Analyysiyksiköiden valintaa ohjasi Tuomen ja Sarajärven (2018, s. 122) ohjeistuksen mukaan tutkimustehtävä ja edelleen tämän työn tutkimuskysymys. Yksiköiksi pyrittiin valikoimaan tutkimuskysymystä mahdollisimman kuvaavia yksittäisiä termejä, joita todennäköisesti tutkimukseen kerätystä aineistosta löytyisi kattavasti. Analyysiyksikön määrittelyssä ei haluttu tehdä liian tiukkaa rajanvetoa ainoastaan edellä lueteltuihin yksittäisiin termeihin. Analyysiyksiköiksi valittiin myös termien ympärille asemoituvat ajatuskokonaisuudet. Määrittely tehtiin siten, että ajatuskokonaisuus voi sisältää yhden tai useampia kokonaisia lauseita, jotta haluttu ajatus ja ilmiö kyetään nostamaan selkeästi ja täysimääräisesti analysoitavaksi. Tärkeäksi asiaksi tämän suhteen katsottiin myös, että ilmiöt ja ajatukset saadaan poimittua mukaan ilman, että niiden merkitys muuttuu tai sitä on vaikea tulkita sen vuoksi, että dataa karsittaisiin liiaksi pois.

Aineistolähtöisen sisällönanalyysin ensimmäinen askel on siis kerätyn aineiston redusointi eli pelkistäminen. Analysoitava data on tässä tapauksessa artikkeleiden ja litteroidun haastattelun tekstiä. Redusointivaiheessa aineistosta pyritään löytämään tutkimuskysymystä vastaava data ja samalla karsitaan epäolennainen pois. Pelkistämisvaiheessa kirjallisesta aineistosta etsitään tutkimustehtävää hyvin kuvailevia ilmaisuja ja ne kirjataan ylös sellaisenaan, jotta alkuperäisdata ei katoa. Alkuperäisilmaisuista muodostetaan edelleen pelkistettyjä ilmauksia, eli alkuperäiset ilmaisut tiivistetään. Yhdestä lauseesta tai ajatuskokonaisuudesta voidaan tehdä useampiakin löydöksiä eli pelkistettyjä ilmauksia. (Tuomi & Sarajärvi, 2018, ss. 122–124.)

Koko sisällönanalyysi ja kaikki sen vaiheet toteutettiin tässä työssä Microsoft Excel -taulukkolaskentaohjelmaa hyödyntäen. Analyysin jokaiselle kolmelle vaiheelle (redusointi, klusterointi, abstrahointi) luotiin omat taulukot, jotta aineistoa olisi helpompi käsitellä ja analyysi pysyy mahdollisimman selkänä. Excelin avulla reitti ylemmiltä tasoilta alkuperäisdataan on myös helposti jäljitettävissä, mikä lisää myös tutkimuksen luotettavuutta.

Excel-taulukon ylälaitaan listattiin ensin tutkimuskysymykset, jotta ne pysyisivät tutkijan mielessä koko analyysiprosessin ajan. Niiden jatkuvan esillä olon oli tarkoitus myös pitää tutkijan fokus tutkittavissa ilmiöissä. Samalla tavoitteena oli, ettei epäolennaisuuksia eksyisi prosessissa mukaan.

Jokainen aineiston osa, eli artikkelin nimi, tekijät ja julkaisuvuosi kirjattiin myös taulukkoon. Tämän johdosta alkuperäiseen aineistoon on helpompi palata tarvittaessa jälkikäteen, koska sellaisia tilanteita hyvin todennäköisesti tulee eteen. Selkeä nimeäminen sekä dokumenttien metatiedot edistävät myös analyysin selkeää struktuuria ja täten koko tutkimuksen luotettavuutta.

Alkuperäisilmaukset kirjattiin omiin soluihinsa ensimmäiseen sarakkeeseen ja pelkistetyt ilmaukset omiinsa toiseen sarakkeeseen. Alkuperäisilmauk-

set, kuten kaikki muutkin analyysin osat, ovat suomennettu. Alkuperäiset englanninkieliset ilmaisut on otettu myös talteen ja ne on kirjattu solukohtaisesti käyttäen Excelin note-ominaisuutta.

Tutkimusta varten toteutettu asiantuntijahaastattelu kävi saman analyysiprosessin läpi kuin kirjallinen aineisto. Litteroitua haastattelua käsiteltiin täysin saman kaavan mukaan ja se tehtiin samassa Excel-taulukossa.

TAULUKKO 2 Sisällönanalyysin ensimmäinen vaihe eli redusointi.

Tutkimuskysymys	
1. Miten selittäviä tekoälymalleja voitaisiin hyödyntää kyberuhkien havaitsemisessa?	
2. Voidaanko selitettävillä tekoälymalleilla saavuttaa korkeampaa luotettavuutta kyberturvallisuuden alalla?	
3. Kuinka selitettävät mallit soveltuisivat organisaatioille, jotka vaativat korkeaa laatua ja tarkkuutta käytettäviltä tietojärjestelmiltä?	
Aineisto/dokumentti	Tekijät
Phishing Email Detection Using Persuasion Cues Phishing-sähköpostin havaitseminen suostutteluvihjeiden avulla	Rohit Valecha , Pranali Mandaokar, and H. Raghav Rao
Tyyppi	Vuosi
Tieteellinen artikkeli	2022
Alkuperäisilmaukset	Pelkistetyt ilmaukset
Tutkimuskysymyksiin vastaamiseksi luomme kolme koneoppimismallia, joissa on relevantteja suostutteluvihjeitä (gain persuasion cues), tappiovihjeitä (loss persuasion cues) ja yhdistettyjä voitto- ja tappiovihjeitä (gain and loss persuasion cues), ja vertaamme estimaatteja perusmalliin, jossa ei oteta huomioon suostutteluvihjeitä. Tulokset osoittavat, että kolme phishing-havaintomallia, joissa on relevantteja suostutteluvihjeitä, ovat F-pistemäärältään noin 5-20 prosenttia paremmat kuin perusmalli, joten ne ovat luotettavia menetelmiä phishing-sähköpostin havaitsemiseen.	- Käytetyn havaitsemismallin luottamuksen kasvattaminen - Suostutteluvihjeiden käyttö phishing-sähköpostien havaitsemisessa
Tällainen tutkimus on hyödyllistä, koska suostutteluvihjeiden syvempi ymmärtäminen voi auttaa suunnittelemaan tehokkaita vastatoimia phishing-sähköpostien havaitsemiseksi ja estämiseksi.	- Suostutteluvihjeiden tehokkuus vastatoimien kehittämisessä
Ongelma kärjistyy entisestään, kun tietojenkalastelijat vaikuttavat uhrien reaktioihin ja pitävät samalla yllä aitoa ja laillista ulkoasua, jolloin suodattimien on vaikea luokitella tietojenkalastelusisältöä vilpilliseksi ja käyttäjät ovat edelleen alttiita tietojenkalasteluviesteille. Näin ollen sähköpostiviestien merkitseminen phishing-sisällöksi on yksi tehtävä, jota monet tietoturva-analyytikot tietoturvaoperaatiokeskuksissa joutuvat tekemään manuaalisesti. Siksi tarvitaan automaattisia, selitettäviä lähestymistapoja, jotka voivat tarjota arvokkaita tietoja phishingin havaitsemiseen sekä phishing-tietojen laadulliseen ja määrälliseen analysointiin.	- Manuaalisen ihimistyön kuormittavuus - Tarve selitettäville lähestymistavoille
Tätä tarkoitusta varten tässä asiakirjassa tarjotaan toimintakelpoista tiedustelutietoa kehittyvistä phishing-hyökkäyksistä (eli kyberuhkien tiedustelutietoa) käyttämällä selitettävyyttä, joka perustuu yhteiskuntatieteellisestä ja psykologisesta kirjallisuudesta peräisin oleviin teoreettisiin näkökulmiin, phishing-hyökkäysten tehokasta havaitsemista varten. Siinä tutkitaan erityisesti suostutteluvihjeiden tehokkuutta phishing-sähköpostin havaitsemisessa tai phishing-sähköpostisuodattimien suunnittelussa.	- Selitettävyyden hyödyt phishingiin liittyvässä kyberuhkatiedustelussa
Tutkijat ovat hiljattain tutkineet suostutteluvihjeiden roolia phishing-alttiudessa ja uhrien käyttäytymisessä. Laajennamme tätä tutkimusta mittaamalla suostutteluvihjeiden tehokkuutta phishing-sähköpostin havaitsemisessa.	- Suostutteluvihjeet selitettävyyttä tukemassa
Tällainen tutkimus on tärkeää, koska syvällisempi ymmärrys phishing-viestintekijöiden taktiikoista voi antaa tietoa sellaisten tehokkaiden vastatoimien suunnittelusta, joilla puututaan suoraan turvallisuusongelmiin ja jotka voivat auttaa phishing-viestien havaitsemisessa ja estämisessä. Tämä tutkimus voi auttaa ymmärtämään tietojenkalasteluviestien havaitsemisjärjestelmää laskennallisten ja inhimillisten tekijöiden avulla. Teoreettisen näkökulman hyödyntäminen voi auttaa tuottamaan selitettäviä malleja, jotka voivat tarjota tulkintoja mustan laatikon malleista (NISTIR 83123).	- Vastatoimien kehittäminen - Ymmärrys hyökkääjän taktiikoista - Teoreettisen näkökulman tuoma hyöty selitettävien mallien tuottamisessa - Black-box -mallien tulkinta
Viimeaikaisessa kirjallisuudessa on kuitenkin tuotu esiin NLP:n haittapuolia phishing-sähköpostin havaitsemisessa, sillä synonyymien käyttöä ja lauserakennetta on vaikea löytää NLP:n avulla. Siinä on myös todettu luokittelijaluokan ongelmia yleisesti, koska koneoppimismenetelmät perustuvat pääasiassa sähköposteissa edustettujen piirteiden tuottamiseen, mikä voi vaatia raskasta manuaalista työtä ja alaan liittyvää asiantuntemusta	- Perinteisten ML-mallien vaatima manuaalinen työ - Vanhat mallit raskaita
Vaikka monissa tutkimuksissa on analysoitu phishing-sähköpostiviesteissä esiintyviä suostutteluvihjeitä ja joissakin tutkimuksissa on osoitettu, että suostuttelu voidaan havaita automaattisesti, on vain rajoitetusti tutkittu suostutteluvihjeiden tehokkuutta phishingin havaitsemisessa. Tämän ongelman ratkaisemiseksi luomme integroidun kehyksen, jonka avulla voidaan kehittää tietokoneella mitattavia ominaisuuksia, jotka kuvaavat voittoa (vastavuoroisuuden, johdonmukaisuuden ja miellyttävyyden kautta) ja tappiota (tappion, vakavuuden ja välittömyyden kautta), jotta phishing-sähköpostiviestejä voidaan havaita.	- Käytetyn havaitsemismallin adaptoituminen ihmismäiseen käytökseen

Sisällönanalyysin toinen vaihe on klusterointi eli ryhmittely. Ensimmäisessä vaiheessa koodatuille alkuperäisilmaisuille suoritetaan tarkka läpikäynti. Tässä vaiheessa aineistosta pyritään löytämään käsitteitä, jotka kuvaavat samankaltaisuuksia. Etsinnän kohteena voivat olla myös eroavaisuuksia kuvaavat käsitteet. Seuraavaksi muodostetaan alaluokat, joiden alle asettuvat luokan nimeen yhdistyvät pelkistetyt ilmaukset. Luokittelun tavoitteena on, että kerättyä tutkimusaineistoa saadaan tiivistettyä. Luokittelu etenee siten, että alaluokkia yhdistellään ja niistä muodostuu yläluokkia. Yläluokkia yhdistelemällä taas muodostetaan pääluokkia ja näiden nimeäminen rakentuu aineistosta nousevien ilmiöiden kuvaavien aiheiden pohjalta. Lopuksi muodostetaan yhdistävät luo-

kat, jotka kytkeytyvät tiiviisti tutkimustehtävään. (Tuomi & Sarajärvi, 2018, s. 124–125.)

Kuten aiemmin mainittua, jokainen analyysiprosessin vaihe kirjattiin omaan taulukkoonsa. Alaluokkien alle koottiin kaikki samankaltaiset pelkistetyt ilmaukset. Luokkia pyrittiin luomaan järkevä määrä, jotta aineiston tiivistäminen pysyisi selkeänä. Luokkien määrän määrittäminen osoittautui kuitenkin haastavaksi.

TAULUKKO 3 Klusterointi.

Pelkistetyt ilmaukset	Alaluokka
Käytetyn havaitsemismallin luottamuksen kasvattaminen Selitettävän havainnointimallin luotettavuus Yleistettävyys Mallin luotettavuus Selitettävyyden varmistaminen Perustelut XAI:n luotettavuudelle Selitettävyyden tärkeys kriittisen informaation käsittelyssä Ennustustulosten luotettavuus Luotettamuksen ja läpinäkyvyyden parantaminen selitysten avulla Analyytikoiden rohkaisu malleihin Selitettävyyden merkityksen korostuminen, kun kyseessä kriittinen infra tai ihmisen henki/terveys Tietoturvaratkaisun toiminnan tärkeys Luottamuksen merkitys tietoturvaratkaisuja kohtaan Tulkittavuuden vaikutus käyttäjien valistukseen Selitettävyyden puutoksen kytkös luotettavuuteen ja uskottavuuteen DL-mallin päätöksenteon tulkittavuus Uskottavuuden lisääminen Selitettävyyden kytkeytyminen analyytikoiden luottamukseen Selitysten tarjoaminen ja luotettavuuden lisääminen AI-menetelmien merkitys APT:n havaitsemisessa Ymmärrys, miksi jokin malli päätyy johonkin ennusteeseen Selittävän mallin tarjoama läpinäkyvyys ja luotettavuus Selitettävyyden vaikutus mallin luotettavuuteen ML-järjestelmien luottamuksen ja oikeudenmukaisuuden varmistaminen Luottamuksen lisääntyminen XAI-mallien suosion kasvaminen Turvallisuuden kehittyminen selitettävyyden avulla Puolustusstrategian kehittäminen AI mallin ymmärrettävyyden vaikutus proaktiiviseen oppimiskykyyn ja havainnointiin	Luottamuksen vahvistaminen
Johdonmukainen suorituskyky luokittelussa Selitettävän havainnointimallin suorituskyky Selitettävän havainnointimallin tehokkuus Suorituskyvyn ja selitettävyyden suhde XAI-mallien potentiaali havaitsemisen parantamisessa Kestävyyden ja luotettavuuden kriittisyys korkean riskin ympäristöissä ML-perustaisten IDS-järjestelmien kehittyminen Selitettävän mallin suorituskyky XAI käsitteen historia XAI:n vaikutus tarkkuuteen Automaattisen uhkatiedon tuottaminen XAI sovellusten käyttö yleisesti XAI menetelmien jakautuminen kahteen kategoriaan LIME:en pohjautuvan XAI-puolustustekniikan tehokkuus XAI menetelmien integrointi CTI alustoihin CTI:n vaatimukset XAI:n tuottamat ennustukset ja selitykset XAI:n apu CTI:n analysoinnissa XAI-mallin tarjoama ennakoiva ja kestävä puolustus, tarkkuus ja selitettävyys Jatkuva oppiminen ja itsenäinen ennakoiva puolustus APT:ta vastaan XAI:n kyky havaita bias Yksittäisen ennusteen selittäminen XAI-mallien kehittynyt menettely XAI:n tehokkuus hyökkäyksien tunnistamisessa XAI-mallin suorituskykyisyys Shapley-arvojen tuoma hyöty tietoverkkoriskien selvittämiseen	Tehokkuus

Kolmas eli viimeinen vaihe on aineiston abstrahointi, jolla tarkoitetaan käsitteellistämistä. Tässä vaiheessa erotellaan oleellinen tieto, joka liittyy tutkimukseen. Tämän valikoidun tiedon pohjalta pyritään muodostamaan teoreettisia käsitteitä. Sisällönanalyysin vaiheet menevät osittain päällekkäin ja katsotaankin, että klusterointi on osa abstrahoinnin prosessia. Abstrahointiin kuuluu eteneminen alkuperäisdatassa ilmaantuvista kielellisistä ilmaisuista käsitteisiin sekä johtopäätöksiin. Abstrahointi on edelleen luokkien yhdistämistä ja sitä to-

teutetaan niin pitkään, kuin se on mahdollista aineiston näkökulmasta. (Tuomi & Sarajärvi, 2018, s. 125.)

Abstrahoinnissa jatkettiin prosessia ja työtä saman kaavan mukaan yhdistelemällä alaluokista yläluokkia. Viimeisenä muodostetiin yhdistävät luokat, joiden tarkoituksena on kuvata tutkittavaa ilmiötä mahdollisimman hyvin ja kiteytetysti.

TAULUKKO 4 Abstrahointi.

Alaluokka	Yläluokka	
Luottamuksen vahvistaminen Tehokkuus Päätöksenteon tukeminen XAI-mallien ja menetelmien hyödyt XAI:n ja kynberuhkatideustelun yhdistäminen XAI-menetelmien moninaisuus	XAI:n tuomat edut ja mahdollisuudet kyberturvallisuuden kentälle	
Selitysten rajoitukset XAI-mallien haasteet XAI:n haavoittuvuudet Tekoälyn implementoinnin haasteet Hyökkääjien käyttämä tekoäly	XAI-mallien keskeneräisyys ja siihen liittyvät haasteet	
Vaikea ymmärrettävyys Läpinäkyvyyden puute	Perinteisten AI/ML-mallien avoimuuden ja tulkittavuuden puuttuminen	
Toimintaympäristön muutos Muuttuvat uhat ja hyökkäysvektorit Selitettävyyden kehitys	Jatkuvasti toimintaympäristö ja kyberturvallisuuskentän muutostila ja siihen vastaamisen tarve	
XAI mallien tarve Tekoälyn rooli Tekoälyn ja ihmisen välinen vuorovaikutus	XAI:n paikka ja rooli kyberturvallisuuden ratkaisuissa	
Tekoälymallien tarve kyberpuolustuskessa Tutkimuksen tarve Tuktimushaasteet	Tutkimuksen puute ja siihen liittyvät haasteet	
Alaluokka	**Yläluokka**	**Yhdistävä luokka**
Luottamuksen vahvistaminen Tehokkuus Päätöksenteon tukeminen XAI-mallien ja menetelmien hyödyt XAI:n ja kynberuhkatideustelun yhdistäminen XAI-menetelmien moninaisuus	XAI:n tuomat edut ja mahdollisuudet kyberturvallisuuden kentälle	**XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa**
XAI mallien tarve Tekoälyn rooli Tekoälyn ja ihmisen välinen vuorovaikutus	XAI:n paikka ja rooli kyberturvallisuuden ratkaisuissa	
Vaikea ymmärrettävyys Läpinäkyvyyden puute	Perinteisten AI/ML-mallien avoimuuden ja tulkittavuuden puuttuminen	**Perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan**
Toimintaympäristön muutos Muuttuvat uhat ja hyökkäysvektorit Selitettävyyden kehitys	Jatkuvasti toimintaympäristö ja kyberturvallisuuskentän muutostila ja siihen vastaamisen tarve	
Tekoälymallien tarve kyberpuolustuskessa Tutkimuksen tarve Tuktimushaasteet	Tutkimuksen puute ja siihen liittyvät haasteet	**XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla**

Tiivistetysti aineistolähtöinen sisällönanalyysi siis koostuu käsitteiden yhdistelemisestä ja siitä voidaan johtaa vastaus tutkimuksen tehtävään ja kysymyksiin. Sisällönanalyysin perustana on tukeutuminen tulkintaan ja päättelyyn. Prosessissa edetään empiirisestä aineistosta käsitteellisempään näkemykseen ilmiöstä, jota tutkitaan. Mitä tulee tuloksiin, pyritään siellä kuvaamaan käsitteet, jotka on luotu luokittelujen pohjalta. Tutkimuksen johtopäätöksissä ja niiden muodostamisessa tutkijan tehtävä on yrittää ymmärtää, mitkä asiat ovat merkityksellisiä tutkittaville. (Tuomi & Sarajärvi, 2018, s. 125.)

3 KYBERTURVALLISUUDEN TOIMINTAYMPÄRISTÖ

Tässä luvussa kuvataan tutkimuksen kohteena olevaa kybertoimintaympäristöä. Ensimmäisessä alaluvussa käsitellään kyberturvallisuutta ylemmällä tasolla, jotta saadaan kattava kokonaiskuva toimintaympäristöstä ja sen sisälle asemoituvista uhkista. Toisessa alaluvussa käsitellään kybertoimintaympäristöä ja kolmannessa keskitytään uhkiin ja niiden nopeaan muutokseen kybertoimintaympäristön sisällä. Lisäksi analysoidaan myös uhkia erilaisten organisaatioiden näkökulmasta sekä miten niihin on varauduttu. Kyberuhkien kehittymistä tarkastellaan samassa luvussa.

3.1 Kybermaailma

Kybermaailmassa toimintaympäristön turvallisuudesta ja suojaamisesta puhuttaessa käsitteitä ja kuvauksia on monia erilaisia. Huomioitavaa käsitteiden käytössä on, mikä on kirjallisuuden sekä käsitteen määrittelijän maantieteellinen alkuperä. Myös siviili- ja sotilasmaailmassa käytetään hieman eri termistöä ja kyberturvallisuuteen nivoutuvia käsitteitä kuvataan hieman eri tavoilla lähestymistavan mukaan. Kyseisessä alaluvussa käsitellään tämän työn kannalta muutama tärkein määrittely. Tässä pro gradu -tutkielmassa nojaudutaan pääasiassa kyberturvallisuuden käsitteeseen.

YK:n määritelmän mukaan kyberillä tarkoitetaan maailmanlaajuista järjestelmää, joka koostuu internetiin liitetyistä tietokoneista, viestintäinfrastruktuureista, verkkoneuvottelujärjestelmistä, tietokannoista sekä tietojärjestelmistä. Tämä kokonaisuus tunnetaan nimellä verkko. Tällä viitataan yleisimmin internetiin. Samaan aikaan termiä voidaan käyttää puhuttaessa jonkin yrityksen, valtion tai muun organisaation tietystä rajatusta sähköisestä tietoympäristöstä. (Andress ym., 2014, s. 4.)

Sanana kyber juontaa juurensa kreikan kielen kantasanasta ”kybereo” ja sillä tarkoitetaan ohjaamista, opastamista ja hallintaa. Kybermaailmaan sisältyy yhden määritelmän mukaan sosio-tekninen ekosysteemi, joka koostuu ihmisistä, datasta, tietoverkoista, tietoteknisistä laitteista sekä ohjelmistoista. Kuvaus on

hyvin samankaltainen kuin perinteinen tietojärjestelmän määritelmä. (Frilander, 2018.)

3.1.1 Kyberavaruus

Ensimmäisenä tehdään katsaus kyberavaruus -käsitteeseen. Stallings (2019, s. 3) kuvaa kirjassaan kyberavaruus-termiä ja tekee siitä havainnollistavan kuvauksen. Sen mukaan kyberavaruuden katsotaan muodostuvan erilaisista artefakteista, joilla on riippuvuussuhde tietokone- ja viestintätekniikkaan tai ne perustuvat näihin. Kyberavaruus muodostuu lisäksi tällaisten artefaktien käyttämästä, tallentamasta, käsittelemästä tai jalostamasta tiedosta sekä yhteyksistä, jotka rakentuvat näiden eri komponenttien välille. (Stallings, 2019, s. 3.)

Kun puhutaan eri osista, joista kyberavaruus koostuu, voidaan nähdä, että ne ovat hyvin alttiita esimerkiksi hakkereiden, rikollisten, terroristien sekä valtiollisten toimijoiden hyökkäyksille sekä vihamielisille toimille. Muutamana esimerkkinä tällaisista toimista voidaan mainita varkaudet, jotka koskettavat arkaluontoisia tietoja, verkossa tapahtuva vandalismi (verkkosivujen turmeleminen (engl. defacement)) sekä palvelunestohyökkäykset. Palvelunestohyökkäykset eri organisaatioita kohtaan ovat varsinkin tänä maailman aikana niin arkipäiväisiä, että niiden voidaan katsoa kuuluvan hiljalleen jopa kansalaisten yleistiedoksi. Tällaisista ikävistä toimista kärsivät niin julkisen kuin yksityisenkin sektorin toimijat aina pienistä organisaatioista suurimpiin mahdollisiin. Nykyään myös eri valtioiden kriittinen infrastruktuuri on hyvin riippuvainen tietotekniikasta ja tietoverkkojen toiminnasta. Kriittiseen infraan voidaan katsoa lukeutuvan esimerkiksi sähköverkot, lennonjohtojärjestelmät, rahoitusjärjestelmät sekä viestintäverkot. Yhtenä murroskohtana tietotekniikan haavoittuvuuksien merkittävyydelle katsotaan olevan syyskuun 11. päivän terrori-iskut New Yorkissa. Sen jälkeen esimerkiksi verkossa tapahtuva vakoilu lisääntyi huomattavasti Yhdysvalloissa yrityksiä sekä valtion virastoja kohtaan. (Clark ym., 2014, s. vii.)

Kriittinen infrastruktuuri on määritelty Suomen huoltovarmuuskeskuksen toimesta sellaisiksi perusrakenteiksi, palveluiksi sekä niihin kytköksissä oleviksi toiminteiksi, jotka ovat välttämättömiä yhteiskunnan elintärkeiden toimintojen ylläpitämiseksi. On tärkeää huomioida, että yksityinen sektori vastaa suurimmasta osasta Suomen kriittistä infrastruktuuria. Siihen lukeutuu sekä fyysisiä laitoksia ja niihin liittyviä rakenteita että digitaalisia palveluita ja toimintoja. Jokainen organisaatio ja yritys, joka lukeutuu tällaiseksi, vastaa itse oman infrastruktuurinsa suojelemisesta. Organisaatiot kuitenkin voivat halutessaan ulkoistaa kriittisen infrastruktuurin, kuten myös palveluiden toteutuksen, jotka liittyvät siihen. Merkille pantava seikka on samaan aikaan, että vastuuta ei voi missään tapauksessa ulkoistaa kenellekään muulle. Myös laki ottaa kantaa joihinkin kriittisiin infrastruktuureihin lakisääteisten velvoitteiden muodossa. (*Ajankohtaisia kysymyksiä ja vastauksia kriittisestä infrastruktuurista ja varautumisesta - Huoltovarmuuskeskus*, ei pvm.)

Myös Averbuch toteaa (2022, s. v) kyberavaruuden muotoutumisesta, että esimerkiksi globaali talous on suurilta osin siirtynyt kyberavaruuteen. Tämän

laajan muutoksen takia tietoverkot ja -virrat, joissa kulkee esimerkiksi finanssialan kannalta merkittävää dataa, ovat prioriteettilistassa korkealla suojaamisen ja varmistamisen suhteen. Tänä päivänä monet verkot nojautuvatkin siihen, että kyberturvallisuuden ylläpitämisen taustalla on haavoittuvuuksien etsiminen sekä löytäminen ja samalla korjausten tekeminen aina, kun heikkoja kohtia löydetään. (Lehto & Neittaanmäki, 2022, s. v.)

3.1.2 Kyberturvallisuus

Koko kybermaailma on hyvin abstrakti ja samaan aikaan tutkijoilla on kyberturvallisuuden määritelmästä useita eri versioita. Sarker ym. (2021, s. 3) nostavat esiin kolme eri määritelmää kyberturvallisuudelle. Ensimmäisenä on määritelmä, jonka mukaan kyberturvallisuus koostuu erilaisista toimista tai politiikoista, jotka liittyvät tietojärjestelmien suojaamiseen uhkilta ja hyökkäyksiltä (Sarker ym., 2021, s. 3).

Toinen määritelmä taas sisältää joukon työkaluja, käytäntöjä sekä ohjeita, joita hyödyntämällä kyetään suojaamaan tietoverkkoja, ohjelmistoja sekä tärkeitä tietoja. Tämän määritelmän mukaan edellä mainittuja artefakteja suojataan hyökkäyksiltä, vahingoittumiselta tai luvattomalta käytöltä. (Sarker ym., 2021, s. 3.)

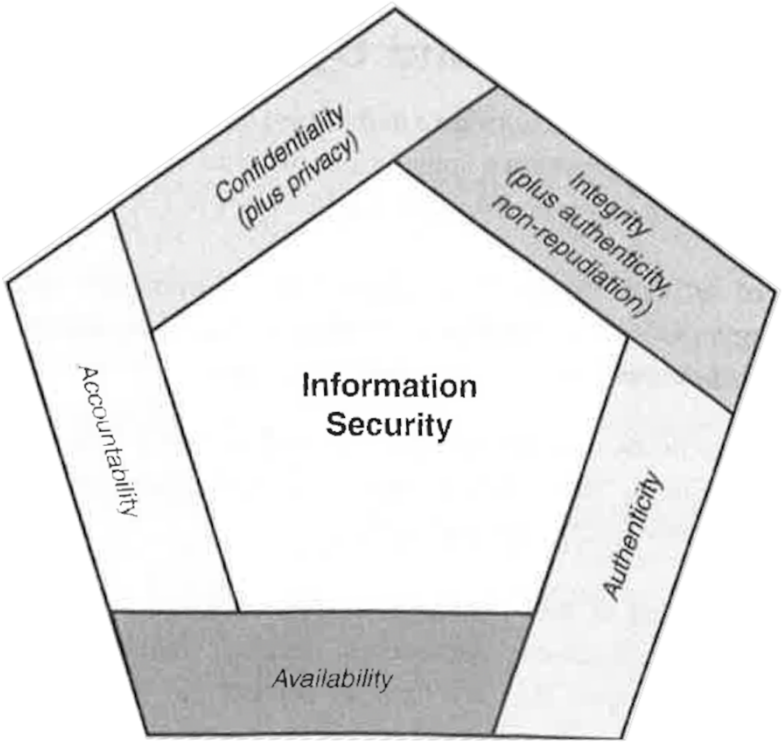
Kolmas määritelmä kuvailee kyberturvallisuutta joukoksi tekniikoita ja menetelmiä, jotka suojaavat tietokoneita, verkkoja, ohjelmia ja tietoja vihamielisiltä toimilta. Suojausta tehdään lisäksi tietojen luvatonta käyttöä, muuttamista ja tuhoamista vastaan. (Sarker ym., 2021, s. 3.) Tästä määritelmästä voidaankin vetää suoraan yhteys seuraavaksi käsiteltävään CIA-kolmioon.

Kyberturvallisuuden voidaan myös katsoa koostuvan kokoelmasta, johon lukeutuu välineitä, politiikkoja, turvallisuusperiaatteita, ohjeita, riskienhallintamenetelmiä, toimia, koulutusta, parhaita käytänteitä, varmennuksia ja teknologioita. Näihin tukeutuen yritetään suojata kyberavaruutta, organisaatioita sekä ennen kaikkea käyttäjiä, jotka hyödyntävät tietotekniikkaa sekä tietoverkkoja. Kyberturvallisuuden tavoitteena on varmistaa organisaation sekä käyttäjien omaisuuden turvallisuusominaisuuksien saavuttaminen sekä ylläpitäminen turvallisuuteen liittyviä riskejä vastaan. Riskit liittyvät kyberavaruuden sisällä oleviin ympäristöihin. (Stallings, 2019, s. 3–4).

Yleisesti kyberturvallisuuden viitekehyksessä turvallisuustavoitteet käsittävät kolme kulmaa ja puhutaan CIA-kolmiosta:

- Tiedon luotettavuus (engl. confidentiality),
- Eheys (engl. integrity),
- Saatavuus (engl. availability).

Tähän kolmioon voidaan ottaa mukaan myös informaation aitous (engl. authenticity) sekä vastuullisuus. Luotettavuuden sisään voidaan lukea tiedon yksityisyys (engl. privacy) ja samalla eheyden sisään lukeutuu tiedon kiistattomuus (engl. non-repudiation). (Stallings, 2019, s. 3–4.) Tietoturva esitettynä seuraavalla sivulla (kuvio 1).



KUVIO 1 Tietoturvan tavoitteet Stallingsin (2019, s. 4) mukaan.

Kyberturvallisuus voidaan jakaa kahteen isompaan kokonaisuuteen, jotka ovat digitaalisessa muodossa olevan tiedon tietoturva sekä verkkoturvallisuus. Tietoturvan alle nivoutuu lisäksi fyysinen turvallisuus, josta konkreettisena esimerkkinä voidaan käyttää paperille kirjoitettua informaatiota. On hyvin yleistä, että kyberturvaa ja tietoturvaa käytetään toistensa synonyymeina ja lomittain. (Stallings, 2019, s. 4.)

Sen sijaan kyberavaruuden sisällä turvallisuus käsitetään tekniikoina, prosesseina ja toimintatapoina, joita hyödyntämällä ja joiden avulla voidaan ehkäistä haitallisia sekä ikäviä tapahtumia. Tällaiset tietojärjestelmiä vastaan toteutetut tapahtumat ovat yleensä vihamielisten ja pahantahtoisten toimijoiden aiheuttamia ja samaan aikaan tahallisia. (D. Clark ym., 2014, s. 2)

Ongelmat, jotka liittyvät kyberturvallisuuteen juontuvat kolmesta eri tekijästä. Ensimmäisenä ongelmatekijänä on ilkeämieliset toimijat ja ryhmät, jotka vaikuttavat kyberavaruuden sisällä. Toisena tekijän on se tosiseikka, että tänä päivänä yhteiskunnat ovat hyvin riippuvaisia tietotekniikasta useiden tärkeiden toimintojen osalta. Kolmas ongelmakohta on haavoittuvuudet, joita tietojärjestelmissä on, ja joita vihamieliset toimijat pyrkivät hyödyntämään. Kyberturvallisuuden tarkoituksena on odotus, jonka mukaan ongelmia aiheuttavista tekijöistä huolimatta tietoteknisten järjestelmien kuuluisi tehdä se, mitä niiden odotetaan tekevän. Samaan aikaan tietotekniikan ei kuuluisi tehdä mitään sellaista, mitä sen ei haluta tekevän. (Clark ym., 2014, s. 2.)

Kyberturvallisuudessa on näiden tosiasioiden valossa kysymys taistelusta, jolla ei tule koskaan olemaan loppua. Nähdään myös, ettei ratkaisua tähän on-

gelmaan ole näköpiirissä. Oman haasteensa kyberturvallisuuteen liittyviin kysymyksiin ja ongelmiin luo nykyaikaisten tietojärjestelmien monimutkaisuus sekä niitä käyttävien ja niiden osana olevien ihmisten tekemät tahalliset tai tahattomat virheet. Lisämausteen kyberturvan rakentamiseen tuo uhat, jotka kehittyvät jatkuvasti ja samaan aikaan uhkatoimijat, jotka käyttöönottavat jatkuvasti uusia työkaluja sekä teknikoita, joita hyödyntämällä ne pyrkivät vaarantamaan turvallisuutta. Tilannetta mutkistaa myös se tosiasia, että tietotekniikan integroituminen yhteiskuntiin yllyttää vihamielisiä toimijoita entisestään toteuttamaan erilaisia pahantahtoisia toimia tietojärjestelmiä kohtaan. Näin ollen voidaan todeta, että kyberturvallisuuden kehittäminen ei ole asia, jonka voisi vain kerralla laittaa kuntoon ja unohtaa sen jälkeen. Kyberturvallisuudessa on kyse ennen kaikkea jatkuvista prosesseista. (Clark ym., 2014, s. 2–3.)

Clark ym. (2014, s. 4) esittävätkin aiheellisen näkökulman ja toteamuksen kyberturvallisuuden kehittämisestä, jonka mukaan ei ole kyse siitä, miten kyberturvallisuus ongelmana voitaisiin ratkaista. Kyse on, miten siitä voidaan tehdä hallittavissa oleva asia. Jotta kyberturvaa horjuttavia rikkomuksia voidaan tehokkaasti torjua, on ymmärrettävä, että yksittäisten henkilöiden, yritysten, valtion virastojen ja koko kansakunnan kyberturvan parantamisella ja kehittämisellä on tässä yhtälössä iso rooli sekä arvo. Onkin tärkeää rakentaa sellainen puolustus, joka jo itsessään vähentää kyberhyökkäyksen tekemistä. Hyvin toteutettu ja kokonaisvaltainen kyberpuolustus ajaa vihamieliset uhkatoimijat tilanteeseen, jossa ne joutuvat käyttämään paljon ajallisia ja taloudellisia resursseja yrittäessään toteuttaa tunkeutumisia jonkin organisaation tietojärjestelmiin. Vankka puolustus niin ikään hidastaa hyökkääjiä, jolloin puolustavalle osapuolelle jää enemmän aikaa reagoida ja torjua hyökkäys. Kyberturvallisuuden parantaminen voidaan sen taustalle asettuvien toimien osalta jakaa kahteen eri kategoriaan. Ensimmäinen sisältää pyrkimykset hyödyntää tehokkaammin ja laaja-alaisemmin asioita, jotka ovat jo tiedossa liittyen turvallisuuden kehittämiseen. Toisen kategorian katsotaan olevan ponnistelut kehittää uutta tietämystä kyberturvallisuuden vahvistamiseksi. (Clark ym., 2014, s. 3.)

3.2 Kybertoimintaympäristön kuvaus

Kybertoimintaympäristöä kuvataan usein erilaisten kerrosarkkitehtuurien avulla. Laari ym. (2019, s. 12–13) jakavat kyberavaruuden kolmeen eri kerrokseen, jotka ovat fyysinen-, looginen- sekä käyttäjäkerros. Fyysinen kerros käsittää kineettisessä maailmassa sijaitsevat kosketeltavissa olevat artefaktit, kuten tietokoneet tai kaapelit. Laitteet, järjestelmät ja infrastruktuurit, jotka lukeutuvat fyysisen kerroksen alle, muodostavat fyysisiä reittejä sekä verkostoja, jonka lisäksi tämä kerros sisältää myös maantieteelliset osat. (Laari ym., 2019, s. 12.)

Loogisen kerroksen artefaktit ovat esimerkiksi erilaisia ohjelmia tai ohjelmakoodia eli asioita, joita ei voi käsin koskettaa. Näin ollen fyysisen maailman maantieteelliset rajat häivyttyvät pois eivätkä loogiseen kerrokseen kuuluvat komponentit tottele fyysisen kerroksen lainalaisuuksia ja sääntöjä. Looginen kerros rakentuu yhteyksistä, jotka asemoituvat verkkojen solmujen välille. Sol-

muilla tarkoitetaan tässä asiayhteydessä alemman kerroksen, eli fyysisen kerroksen laitteita. Tällaisia voivat olla esimerkiksi tietokoneet sekä erilaiset mobiililaitteet. Oleellinen osa solmuja on niiden sisältämät verkkoasetukset, tiedonsiirtoprotokollat, verkkotunnukset, sovellukset ja protokollat ja niin edelleen, jotka näyttelevät merkittävää roolia keskustelussa fyysisen kerroksen kanssa. (Laari ym., 2019, s. 13.)

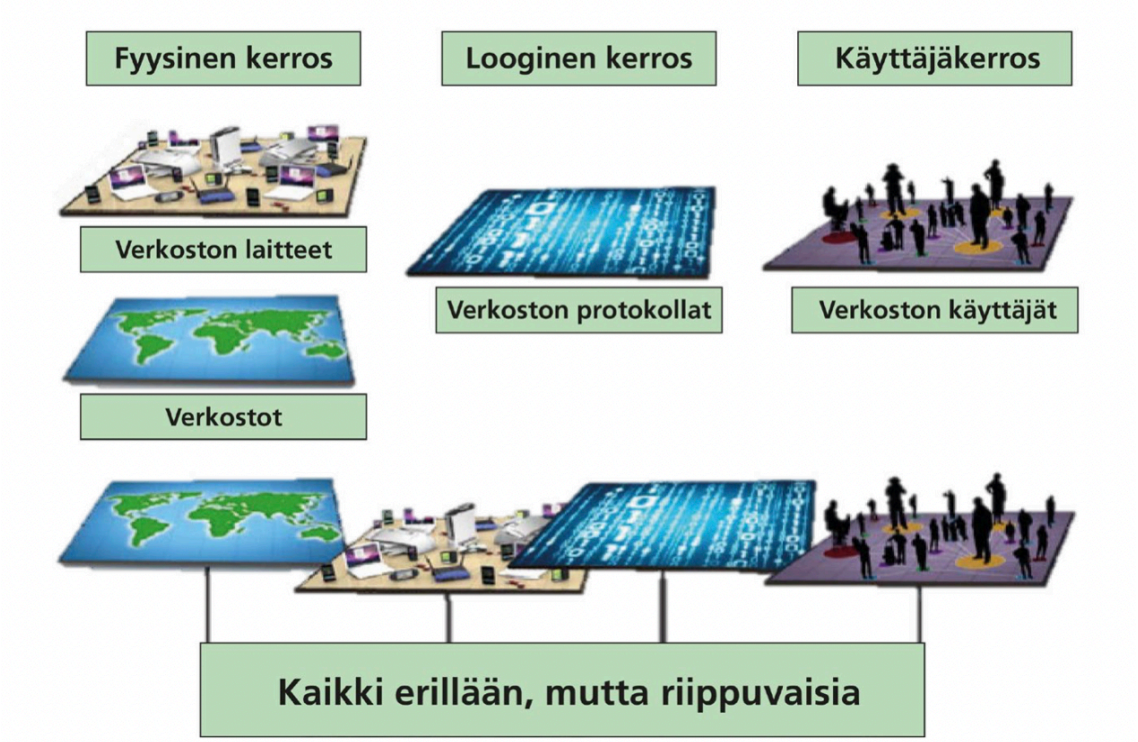
Käyttäjäkerrokseen lukeutuu nimensä mukaisesti ihmiset, jotka käyttävät tietoteknisiä laitteita. Kyseinen kerros rakentuu tekijöistä, jotka liittävät ihmiset osaksi kybertoimintaympäristöä. Tästä esimerkkinä toimii verkossa käytettävät nimimerkit, jotka sittemmin yhdistetään virtuaalisiin osoitteisiin. Kybertoimintaympäristössä pahantahtoisten ihmisten vastuuseen saattaminen on hyvin haastavaa, koska esimerkiksi yhdellä ihmisellä voi olla lukematon määrä eri käyttäjänimiä ja identiteettejä verkossa. Ihminen on tässä viitekehyksessä yleensä se kaikista heikoin lenkki, johon on helpointa vaikuttaa ja sitä kautta päästä tunkeutumaan tietojärjestelmien sisälle ja aiheuttaa siellä tuhoa. Ihmisiltä vaaditaankin tänä päivänä kybermaailmassa huomattava määrä erilaisia taitoja, varovaisuutta sekä harkintaa ja kyberturvallisuus nähdäänkin joukkuepelinä, jossa jokaisella siellä vaikuttavalla ihmisellä on suuri rooli. (Laari ym., 2019, s. 14.)

Suomen ulkoministeriö on määritellyt (*Kyberturvallisuus ja kybertoimintaympäristö*, ei pvm.) kybertoimintaympäristön ihmisten luomaksi digitaaliseksi rinnakkaistodellisuudeksi. Se yhdistää maailmanlaajuisesti informaatioteknologian, automatisoitujen ohjausjärjestelmien, internetin ja sosiaalisen median kautta toisiinsa ihmisiä sekä laitteita riippumatta valtioiden maantieteellisistä rajoista. Toiminnot, jotka ovat jokapäiväisen elämän kannalta elintärkeitä, ovat vahvassa riippuvuussuhteessa tietoverkkoihin. Tällaisiin toimintoihin lukeutuvat esimerkiksi teollisuus, vesi- ja energiahuolto, pankkijärjestelmä, terveydenhuolto sekä liikenne. (*Kyberturvallisuus ja kybertoimintaympäristö*, ei pvm.)

Valtiotasolla modernit yhteiskunnat, kuten Suomi, ovat erittäin haavoittuvaisia tietoverkkojen ja -järjestelmien häiriöille korkean riippuvuuden vuoksi. Suomen valtion toimesta kybertoimintaympäristön herkkyys ja siellä vaanivat uhat on otettu huomioon Suomen kyberturvallisuusstrategiassa. Siinä määritellään keskeiset tavoitteet sekä toiminteet, joita hyödyntämällä kyetään vastaamaan kybertoimintaympäristön haasteisiin ja samalla varmistumaan sen toimivuudesta. (*Kyberturvallisuusstrategia*, ei pvm.)

Vuonna 2019 julkaistu Suomen kyberturvallisuusstrategian mukaan Suomen tavoitteena on huolehtia kybertoimintaympäristöstään nojautuen aktiiviseen kansainväliseen- sekä EU-yhteistyöhön. Strategiassa mainitaan, että kansainvälinen yhteistyö on Suomen kyberturvallisuuden kannalta elintärkeää sekä teknisellä että poliittisella tasolla. Kybertoimintaympäristön suojaaminen tapahtuu toimenpiteillä, jotka nostavat vastapuolen ja vihamielisten uhkatoimijoiden rimaa tehdä kyberhyökkäyksiä Suomea vastaan. Kynnyksen korottaminen toteutetaan esimerkiksi parantamalla kyberhyökkäysten havainnointi – ja attribuutiokykyä sekä kykyä vastatoimiin. Tässä viitekehyksessä vastatoimet voivat olla muun muassa lainvalvontatoimia, diplomaattisia toimenpiteitä tai aktiivisia kybervastatoimia. (*Suomen kyberturvallisuusstrategia 2019 – Turvallisuuskomitea*, ei pvm.)

Kybertoimintaympäristön rakenne kuvattuna (kuvio 2) (Laari ym., 2019, s. 13).



KUVIO 2 Kybertoimintaympäristö Laarin ym. mukaan (2019, s. 13).

Clark (2010, s. 2–4) sekä Klimburg & Mirtl (2012, s. 11–14) puolestaan ovat esittäneet kontekstuaalisen mallin, joka koostuu neljästä kerroksesta. Tätä mallia voidaan hyödyntää rakenteellisena viitekehyksenä kyberavaruudessa.

- Fyysinen kerros kuvastaa kaikkia laitteita, jotka ovat kyberavaruudessa.
- Looginen kerros puolestaan kuvailee loogisia palasia, jotka tukevat kyberavaruuden infrastruktuuria.
- Tietokerrokseen sisältyy kaikki mahdollinen luotu ja tallennettu informaatio sekä raakadata. Kyberyksiköt välittävät ja käsittelevät tätä tietoa ja dataa kyberavaruuden välityksellä.
- Ihmiset-kerros havainnollistaa inhimillistä puolta, joka liittyy kyberavaruuteen. Tämä kerros kommunikoi ja hyödyntää kyberavaruuden toimintoja.

3.3 Uhkat kyberdomainissa

Kyberuhat ovat kiinnostavia siitä syystä, että niitä tulee olemaan ympärillämme niin kauan, kuin on olemassa digitaalisia laitteita, halusimme sitä tai emme. Kuten Hyppönenkin (2022, s. 9) toteaa, ”internet on parasta ja pahinta, mitä meille on tapahtunut.” Hyppösen mukaan (2022, s. 9) digitalisaation vallankumous on ympärillämme ja näkyvissä joka paikassa arkisessa elämässämme. Samaan aikaan internet tuo karmivia riskejä vaikkakin samalla se toki tuottaa myös merkittävästi uusia hyötyjä. Internetiin liittyvät haasteet ovat muuttuneet niin radikaalisti globaaleiksi ongelmiksi, että yhden organisaation tai edes valtion on täysin mahdotonta ratkaista niitä. (Hyppönen, 2022, s. 15.)

Kyberuhka terminä voidaan määritellä uhkaksi, joka kohdistuu digitaalisten teknologioiden kautta. On myös tärkeää huomioida, että suurin osa digitaalisista järjestelmistä sisältävät haavoittuvuuksia. Samalla on kuitenkin ymmärrettävä, ettei kaikkiin haavoittuvaisiin järjestelmiin kohdistu suoraa hyökkäyksen uhkaa. Tavallisille internetin käyttäjille ja kansalaisille todennäköisiä kyberuhkia voisivat olla esimerkiksi kodin internetyhteydessä tapahtuva katkos tai sähköposti- tai sosiaalisen median tiliin kohdistuva tietomurto. Osa tällaisista pienen mittakaavan yksittäisiä henkilöitä koskevista uhkista on enemmänkin kiusallisia, kun taas osa voi aiheuttaa jo merkittävämpääkin haittaa, kuten esimerkiksi henkilökohtaisten tietojen varastaminen. (Laari ym., 2019, s. 28.)

Laaja-alaisempia kybertoimintaympäristöön kohdistuvia hyökkäyksiä voi olla esimerkiksi sähköverkkoihin tai puolustusalan organisaatioiden tietoverkkoihin kohdistuvat vihamieliset toimet. Tällaiset voivat aiheuttaa vakaviakin seurauksia ja pahimmillaan jopa ihmishenkien menetyksiä. (Laari ym., 2019, s. 27.) Laari ym. ovat avanneet muutamia keskeisiä ja yleisiä uhkaan liittyviä käsitteitä teoksessaan (2019, s. 29).

TAULUKKO 5 Yleisiä uhkaan liittyviä käsitteitä (Laari ym., 2019, s. 29).

Uhka	Mahdollisesti toteutuva haitallinen tapahtuma tai kehityskulku. Sotilaallisesti uhka esitetään usein muodossa uhka = kyky x tahto.
Tietoturvauhka	Mahdollisesti toteutuva haitallinen tapahtuma tai kehityskulku, joka kohdistuu tietoturvaan ja toteutuessaan vaarantaa sen.
Kyberuhka	Mahdollisesti toteutuva haitallinen tapahtuma tai kehityskulku, joka kohdistuu kybertoimintaympäristöön ja toteutuessaan vaarantaa siitä riippuvaisen toiminnon. Kyberuhkat voivat aiheutua paitsi toteutuneista tietoturvauhkista myös digitaalisessa viestintäympäristössä toteutettavista, yhteiskunnan turvallisuutta vaarantavista teoista.
Haavoittuvuus	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.
Nollapäivä-haavoittuvuus	Nollapäivähaavoittuvuus on tietojärjestelmässä haavoittuvuus, joka ei ole yleisesti tiedossa ja johon ei ole saatavilla korjausta.
Hakkeri	Henkilö, joka tunkeutuu tai vaikuttaa tietoverkkoon, tietojärjestelmään tai niiden sisältämään tietoon ja käyttää ohjelmaa, palvelua tai muita resursseja.

3.3.1 Uhkatoimijat

Kybermaailmassa omia tavoitteitaan edistävien vihamielisten ja pahantahtoisten toimijoiden, eli uhkatoimijoiden, kirjo on hyvin laaja ja niiden koko vaihtelee yksittäisistä toimijoista aina suuriin valtiollisiin ryhmiin. Uhkatoimijoiden motiivit vaihtelevat laajalti huvin vuoksi tehtävästä ilkivallasta haluun varastaa toiselta. Joidenkin toimijoiden motiivina ja niiden taustatekijöinä voivat olla sen sijaan erilaiset ideologiat tai kansallismieliset ajatukset. Samoin kuin uhkatoimijoiden koko ja motiivit, vaihtelevat myös niiden taidot. On olemassa toimijoita, joiden tekninen osaaminen ja ymmärrys on hyvin alkeellisella tasolla, ja he kykenevät käyttämään vain muiden kehittämiä työkaluja. Samanaikaisesti kaikista kehittyneimmät taidot omaavilla ryhmillä on kykyjä, joilla he kykenevät tunnistamaan kohdejärjestelmien ja -verkkojen heikot kohdat. Tällaiset toimijat pystyvät tietojärjestelmien heikkouksien perusteella kehittämään omia työkaluja. Edistyksellisimmillä ryhmillä on myös kattavat resurssit, joiden takaa löytyy valtioita tai muita suuria organisaatioita kuten järjestäytyneen rikollisuuden järjestöjä. Ryhmät, joita ohjaavat ja rahoittavat kansallisvaltiot, voivat saada käyttöönsä kansallisen tiedustelun, armeijan ja lainvalvontaviranomaisten resursseja. Puolustavan osapuolen kannalta ikävä seikka on, että tavallisten tietoverkkorikollisten varalta toteutetut puolustukselliset toimet ovat vain hidasteita edistyksellisille toimijoille – eivät esteitä. (Clark ym., 2014, s. 49.)

3.3.2 APT-uhat

Kaikista kehittyneimmistä ja ammattimaisimmista uhkatoimijoista käytetään englanninkielistä nimitystä Advanced Persistent Threat (APT). Suomeksi käännettynä puhutaan kehittyneistä ja pysyvistä uhista. APT-termi juontaa juurensa Yhdysvalloista, jossa sitä käytettiin ensimmäisen kerran vuonna 2006 ilmavoimien toimesta. Sillä viitattiin tuolloin kehittyneeseen vastustajaan, joka kykenee harjoittamaan kybersodankäyntiä pitkällä aikavälillä strategisten tavoitteiden tukemiseksi. Voidaan siis todeta, että APT:ta voitaisiin kuvailla toimijaksi, joka on siis teknisesti hyvin kehittynyt ja sitkeä, eli sillä on kyky ja tahto toteuttaa hyökkäyksellisiä toimia pitkäjänteisesti ja kärsivällisesti pitkiäkin ajanjaksoja. APT:lla tarkoitetaan ja kuvataan tämän lisäksi luonnetta, joilla APT-toimija tunkeutuu sen kohteena olevaan tietoverkkoon. Sen toiminta on teknisesti edistynyttä, jonka lisäksi sitä voi olla hyvin haastava jäljittää ja tuhota. Näin ollen sen toiminta on jatkuvaa. APT-toimijat keskittyvät yleensä johonkin tiettyyn tarkoin valikoituun kohteeseen, joka on erityisen arvokas. APT-toimijat käyttävät useimmiten työkaluja sekä tekniikoita, jotka ovat räätälöity hyökättävän kohteen mukaan, jonka lisäksi ne pyrkivät olemaan mahdollisimman huomaamattomia, niin pitkään kuin mahdollista. (Clark ym., 2014, s. 50.)

APT-ryhmille ominaista on myös, että niiden yhtenä tavoitteena on organisaation tietojärjestelmään sisään päästyään pysyä siellä niin kauan, kunnes asetettu päätavoite on saavutettu. APT-hyökkäyksissä on useita eri vaiheita ja useissa tapauksissa toimijoiden pääasiallisina pyrkimyksinä on vakoilu sekä tietojen varastaminen kohdeorganisaatiosta. Sen vuoksi APT-hyökkäyksiä pidetäänkin monimutkaisina sekä monivaiheisina. APT-toimijat ja niiden toteuttamat toimet asettavatkin nykyisille havaitsemismenetelmille ja -järjestelmille suuren haasteen kehittyneiden hyökkäystekniikoiden ja -taktiikoiden johdosta. Ne käyttävät myös tuntemattomia haavoittuvuuksia. Onnistuessaan APT-toimijoiden vihamieliset toimet aiheuttavat sen kohteeksi joutuneelle instanssille merkittäviä taloudellisia vahinkoja ja näin ollen hyökkäyksistä aiheutuneet kustannukset ovatkin yksi syy ja hyvinkin merkittävä motivaattori tunkeutumis- ja havaitsemisjärjestelmien investoinneille. Onkin todettu, että APT-ryhmien toteuttamat hyökkäykset ovat tällä hetkellä hyvin vakava uhka yksityisille yrityksille kuten myös valtioille. (Lemay ym., 2018.)

Huolestuttava seikka APT-ryhmiä vastaan juostavassa kilpajuoksussa on, että paikkansa vahvimmin vakiinnuttaneimmat toimijat ovat osoittaneet korkeaa sopeutumiskykyä ja kehittymistä. Harvoin käy niin, että toimijan paljastuttua se käyttäisi samoja taktiikoita, tekniikoita ja menetelmiä toisen kerran. Sen jälkeen APT-toimijat muuttavat yleensä toimintatapojaan ja puolustava osapuoli onkin usein jäljessä ja myöhässä. Jos kehitys puolustustekniikoiden osalta jatkuu sellaisena, ettei niihin onnistuta löytämään merkittäviä parannuksia, APT-hyökkäyksien torjuntaan liittyvät haasteet tulevat muuttumaan entistä suurempina. Tutkimuskentällä on todettu, että APT-toimijoiden aiheuttamien uhkien torjuntaan kehitettyjen puolustusratkaisujen tutkimuksessa olisi tärkeää kiinnittää huomiota nykyisiin valmiuksiin sekä valmiuksien kehityshistoriaan. Näin voitaisiin varmistaa puolustusratkaisujen osalta, mikä niiden arvo on pitkällä aikavälillä. Tietämys APT-toimijoista on samaan aikaan edelleen puutteellista ja

useimpien kaikista kehittyneimpien toimijoiden toteuttamien hyökkäysten alkuperäiset hyökkäysvektorit ovat hämärän peitossa. Useilla ryhmillä on lisäksi kyky piilottaa omat jälkensä tehokkaasti, joka luonnollisesti asettaa haastetta hyökkäyksien tutkimiselle. Tästä on tehty johtopäätös, jonka mukaan APT-toimijoiden menetelmien tutkiminen olisi tutkimisen arvoista. (Lemay ym., 2018, s. 53.)

4 TEKOÄLY KYBERTURVALLISUUSYMPÄRISTÖSSÄ

Tämä luku käsittelee tekoälyä. Se mullistaa maailmaa parhaillaan enemmän kuin mikään muu teknologian saralla. On esitetty näkemyksiä siitä, kuinka tekoälyä on rinnastettu siihen, kun internet yleistyi 1990- ja 2000-luvuilla ja samalla muutti maailmaa tavalla, jota ei olisi voitu edes kuvitella esimerkiksi 1900-luvun alussa. Ihmiskunnalla on toki ollut suuria ja villejäkin visioita siitä, kuinka autot lentävät ja maailmanloppu tulee milleniumina, eli kun vuosituhat vaihtuu 2000-luvulle. Kukapa olisi osannut 1970-luvulla kuvitella, että miltei jokaisen länsimaissa elävän nuoren ja aikuisen ihmisen taskussa on lähes kaikkitietävä vastauskone? Nyt näitä kutsutaan Googleksi ja ChatGPT:ksi.

Tämän luvun ensimmäisessä alaluvussa käsitellään tekoälyä ensin hieman yleisemmin ja syvennytään sen jälkeen osiin, joista se koostuu. Toinen ja kolmas alaluku käsittelee tekoälyn hyödyntämistä sekä hyökkääjän että puolustajan näkökulmista. Neljännessä alaluvussa syvennytään tutkielman kannalta tärkeimpään ja kiinnostavimpaan aihealueeseen eli selittäviin tekoälymalleihin. Tavoitteena on saada hyvä peruskäsitys aiheesta, joka tukee tutkijan tekemää sisällönanalyysia myöhemmin tässä työssä.

4.1 Mitä on tekoäly?

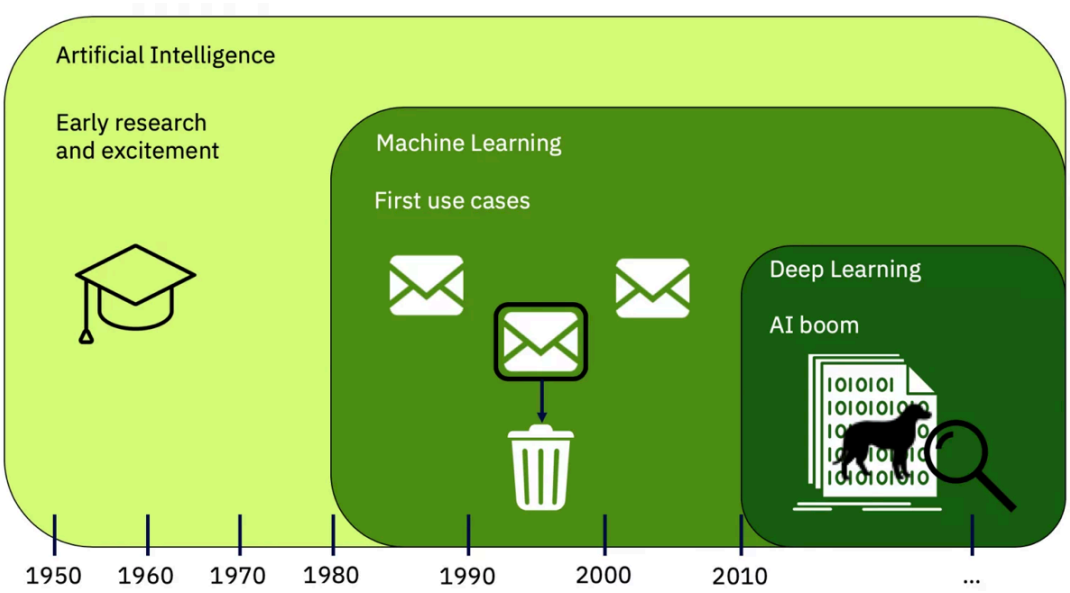
Tekoälystä puhutaan kilvan alasta riippumatta. Sen tuomista mahdollisuuksista keskustellaan monessa käänteessä tänä päivänä. Toki uhkistakin – onneksi. On aiheellista esittääkin kysymys; onko maailma valmis näin nopealle murrokselle, jonka tekoäly mukanaan tuo sekä hyvässä että pahassa? Onkin tärkeää tämän pro gradu -tutkielman kannalta tehdä määritelmä, mikä tai mitä tekoäly oikein on ja mistä se juontaa juurensa. Täten voidaan ymmärtää myöhemmin käsiteltävää selittävää tekoälyä ja sen soveltumista kyberuhkien havaitsemiseen. Tekoälylle ei ole olemassa yhtä yksiselitteistä määritelmää, mutta tässä luvussa

pyritään avaamaan se yläkäsitteenä mahdollisimman selkeästi ja riittävän kattavasti.

Kuten Russel ym. (2022, s. 19) toteavat, tekoälyn sanotaan usein olevan mielenkiintoinen käsite, mutta harvemmin määritellään, mikä tai mitä se on. Historian valossa tutkijat ovat esittäneet erinäisiä määritelmiä siitä, mitä tekoäly heidän mielestään on. He ovatkin pyrkineet löytämään eri versioita tekoälystä. On tutkijoita, jotka ovat rinnastaneet tekoälyn inhimillistä tasoa vastaavalle suorituskyvylle. Samaan aikaan esiintyy määritelmiä, joiden mukaan parempi kuvaus olisi abstrakti ja muodollinen määritelmä, jota voidaan kutsua termillä rationaalinen. Tätä kuvaillaan ”oikean asian tekemiseksi”. Yhtä aikaa rationaalisuuden kohteessa nähdään kuitenkin varianssia. On tahoja, jotka kuvailevat älykkyyttä sisäisten ajatteluprosessien ja päättelyn ominaisuutena. Toiset tutkijat taas keskittyvät älykkääseen käyttäytymiseen. (Russell ym., 2022, s. 19.)

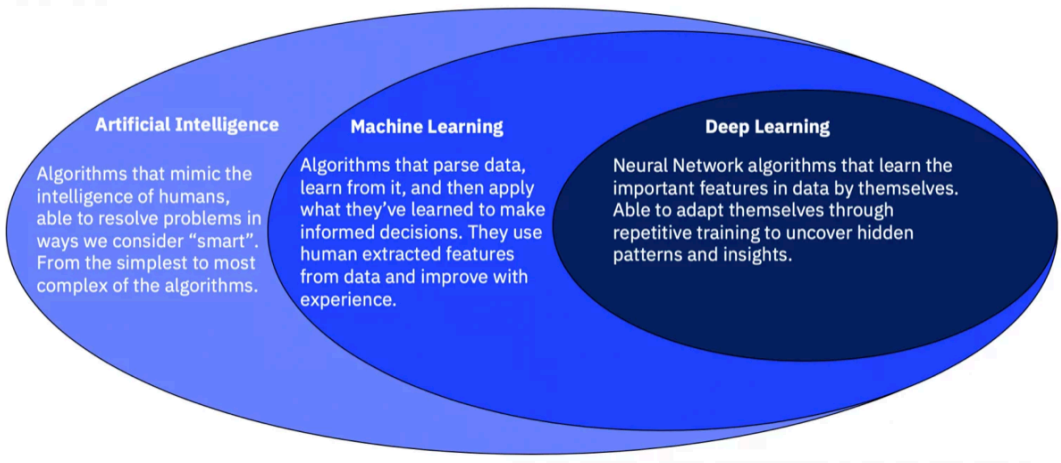
Onkin ilmeistä, että voidaan rakentaa neljä yhdistelmää edellä mainituista ulottuvuuksista, joihin lukeutuu inhimillinen vs. rationaalinen ja ajattelu vs. käyttäytyminen. Tekoälyn määrittelyssä käytetyt menetelmät ovat keskenään erilaisia. Kun puhutaan tavoittelusta, joka tähtää ihmisen kaltaiseen älykkyyteen, on sen väitetty olevan osittain psykologiaan kytkeytyvää empiiristä tiedettä. Siihen liittyy havaintoja sekä hypoteeseja todellisesta ihmisen käyttäytymisestä ja ajatteluprosesseista. Tarkasteltaessa rationaalista lähestymistapaa, siihen sisältyy tekniikan ja matematiikan yhdistelmä. Samanaikaisesti rationaalinen lähestymistapa nivoutuu tilastotieteeseen, kontrolliteoriaan sekä taloustieteeseen. Eri laidoilla olevat ryhmät ja tutkijat ovat sekä väheksyneet että tukeneet toisiaan. (Russell ym., 2022, s. 19–20.)

Garnham (2018, s. 1) puolestaan määrittelee tekoälyn lähestymistavaksi käyttäytymisen ymmärtämisen. Sen pohjana on oletus, että älykkyyttä voitaisiin analysoida parhaiten siten, että yritetään toisintaa ja matkia sitä. Tässä kontekstissa sillä tarkoitetaan simulointia tietokoneen avulla. Tekoälyn katsotaan lukeutuvan osaksi tietojenkäsittelytieteitä. On myös huomionarvoista, että tekoäly on omana tutkimuksen alana verrattain uudehko, sillä sen voidaan sanoa saaneen alkunsa vasta 1950-luvun puolivälissä. (Garnham, 2018, s. 1.)



KUVIO 3 Tekoälyn historia tiivistetysti (Ceron, 2019).

Tekoäly on tutkimusalana pohjimmiltaan käyttäytymisen tutkimista ja yksi sen päätavoitteista on ymmärtää ihmisen älykkyyttä. Toisena tavoitteena tulee hyödyllisten koneiden tuottaminen. Garnham (2018, s. 2) toteaa, että tekoäly terminä on epäonninen, koska molemmat edellä mainitut osat ovat jokseenkin harhaanjohtavia. On kuitenkin samalla hyvä katsoa englanninkielisen termin ensimmäistä osaa, artificial, hieman syvällisemmin, koska se tarkoittaa keinotekoista – ei todellista. (Garnham, 2018, s. 2)



KUVIO 4 Tekoälyalgoritmien erikoistuminen (Ceron, 2019).

Edellä olevissa kappaleissa kuvattuja määritelmiä myötäilee myös Vähäkainun ym. (2018) tutkimus, joka käsittelee tekoälyä kyberturvallisuuden kentällä. Kyseisessä tutkimuksessa tekoälystä esitetään määritelmä, jonka mukaan sitä voitaisiin kuvailla keinotekoisena älykkyytenä. Se mahdollistaa monimutkaisten ongelmien ratkaisun koneen avulla. Määritelmää jatketaan kuvaamalla tekoälyä tietotekniikan ja fysiologisen älykkyyden yhdistelmäksi, joka edesauttaa, että on mahdollista päästä haluttuihin tavoitteisiin laskennallisesti. (Vähäkainu ym., 2018, s. 4.)

Vähäkainu ym. (2018, s. 4) esittävät lisäksi havainnollistavan jaotteluesimerkin, jonka mukaan tekoäly voitaisiin jakaa eri sovellusalueisiin.

- Kognitiivisen tieteen sovellukset,
- Robotiikan sovellukset,
- Luonnollisen kielen sovellukset.

4.1.1 Turingin testi

Russel ym. (2022) jakavat tekoälyn neljään eri lähestymistapaan, joista ensimmäistä avataan seuraavaksi hieman tarkemmin. Ensimmäisenä on siis esitetty Turingin testiin nivoutuvaa lähestymistapaa. Turingin testin kehitti Alan Turing vuonna 1950 ja hän suunnitteli sen ajattelulliseksi kokeeksi, jonka avulla voitaisiin sivuuttaa filosofisesti epämääräiseksi todettu kysymys siitä, ”voiko kone ajatella?”. Tämä kuuluisa testi toimii kaikessa yksinkertaisuudessaan siten, että tietokone läpäisee sen, jos ihminen testin toisena osapuolena ei kykene kirjallisia kysymyksiä esitettyään erottamaan, ovatko vastaukset peräisin ihmiseltä vai tietokoneelta. Testi on edelleen relevantti ja tänäkin päivänä on todettu, että tietokone, joka pystyy läpäisemään testin, vaatii merkittävän määrän kehitystyötä ja vaivannäköä sen ohjelmoimiseksi. Testin läpäisevällä tietokoneella pitää olla ainakin seuraavat ominaisuudet ja kyvyt:

- Kyky luonnollisen kielen käsittelyyn,
- Tiedon esittäminen,
- Automaattinen päättelykyky kysymyksiin vastaamiseksi sekä uusien johtopäätösten tekemiseksi,
- Koneoppimista, jonka avulla se pystyy sopeutumaan muuttuviin olosuhteisiin ja havainnointikykyä, jotta se voi havaita sekä ekstrapoloida malleja. (Russell ym., 2022, s. 20.)

Turing totesi, että älykkyyden osoittamiseksi testiin ei tarvitse sisällyttää ihmisen fyysistä simulointia. Tutkijayhteisössä on ehdotettu myös, että ”täydellinen” Turingin testi olisi sellainen, jonka vaatimuksena olisi vuorovaikutus todellisen maailman esineiden sekä ihmisten kanssa. Tietokoneen, joka suorittaa ja kykenee läpäisemään tällaisen ”täydellisen” Turingin testin, on omattava konenäköä sekä puheentunnistusta siihen, että se kykenee hahmottamaan ympäröivää maailmaa. Lisäksi se tarvitsee robotiikkaa esineiden käsittelyyn ja liikuttamiseen. (Russell ym., 2022, s. 20.)

Russelin ym. (2022, s. 20) mukaan nämä alkuperäisen Turingin testin neljä ominaisuutta ja myöhempien tutkijoiden esittämän niin kutsutun ”täydellisen” tai ”totaalisen” Turingin testin kaksi lisäominaisuutta muodostavat leijonan osan tekoälystä. Testin läpi pääsemisen eteen ei olla kuitenkaan osoitettu kovin suuria ponnisteluita sen vuoksi, että on merkityksellisempää tutkia niitä periaatteita ja fundamentteja, jotka ovat älykkyyden taustalla. (Russell ym., 2022, s. 20.) Muita tekoälyn tutkimisen lähestymistapoja esitetään olevan:

- Inhimilliseen ajatteluun nojautuva näkökulma, jota kutsutaan kognitiivisen mallintamisen lähestymistavaksi.
- Rationaalisen ajattelun lähestymistapa, jota kutsutaan myös ”ajattelun lait” -lähestymistavaksi.
- Rationaalinen toiminta eli rationaalisen toiminnan lähestymistapa, jolla viitataan rationaalisiin agentteihin. (Russell ym., 2022, s. 20–21.)

4.1.2 Määrittelyn monet kasvot

On selvää, että kysymys ”mitä tekoäly on?”, on helppo esittää, mutta sille yksiselitteisen vastauksen löytäminen onkin haastavampi tehtävä. Ei ole olemassa yksimielistä määritelmää sille, mitä tekoäly on. Samalla on väitetty, ettei koneen tuottamalla älyllä olisi vielä toistaiseksi paljoakaan yhtäläisyyksiä ihmisen älykkyyden kanssa. Historian saatossa tekoälylle on muodostunut lukuisia määritelmiä, joista jokainen lähestyy asiaa hieman erilaisista näkökulmista. Monissa lähestymistavoissa on kuitenkin ajatus siitä, että tekoälyllä tarkoitetaan ohjelmistoja tai koneita, jotka kykenevät käyttäytymään ihmisen kaltaisesti. Ajatus ihmisen kaltaisesta älykkyydestä on samaan aikaan ongelmallinen ja haasteellinen, koska ihmisen älykkyyden määritteleminen tai sen mittaaminen on vaikeaa. Monille kulttuureille on lisäksi ominaista asioiden pelkistys numeerisiin mittoihin, jolloin vertailu on luonnollisesti helpompaa. Esiin nousee tällaisessa pelkistämisessä ongelma, jossa kuva objektiivisuudesta voi olla virheellinen ja harhaanjohtava. (Kaplan, 2016, s. 1–2.)

Kun ihmisen kykyjä käytetään tekoälyn mittarina, on se ongelmallista myös siitä näkökulmasta, että koneilla on kyky suorittaa sellaisia tehtäviä, joihin ihminen ei kykene lainkaan. Voi tulla tunne ja vinoutuma siitä, että tämänkaltaiset suoritukset ilmenisivät osoituksena älykkyydestä. Konkreettisia esimerkkinä tällaisesta ilmiöistä voisi käyttää tietoturvaohjelmistoa, joka voi epäillä verkkohyökkäystä sen perusteella, että viidensadan millisekunnin aikana verkkoliikenteessä havaitaan epätavallisia tietopyyntöjä tietyn kaavan mukaan. (Kaplan, 2016, s. 4.)

Toisena esimerkkinä voidaan käyttää tilannetta, jossa tsunamivaroitusjärjestelmä antaa hälytyksen havaitessaan merenpinnan korkeudessa muutoksia, joita tuskin huomaa. Samaan aikaan ne kielivät monimutkaisista merenalla tapahtuvista maantieteellisistä muutoksista. Tällaiset järjestelmät ovat yleistyneet ja tulevat lähitulevaisuudessa yleistymään yhä enenevissä määrin, mutta niiden käyttäytymistä ei voi verrata ihmisen kykyihin. (Kaplan, 2016, s. 4.)

Tästä ristiriidasta huolimatta on todennäköistä, että tällaisia järjestelmiä tullaan pitämään keinotekoisesti älykkäinä. Älykkyyden tunnusmerkkinä pidetään myös sitä, kuinka epäonnistumme, sillä myös älykkäät koneet tekevät virheitä. Asiantuntijuuden tunnusmerkistöön kuuluukin omien rajojen ymmärtäminen ja niiden kunnioittaminen sekä inhimillisten virheiden tekeminen. (Kaplan, 2016, s. 4.)

4.1.3 Koneoppiminen

Koneoppiminen (engl. machine learning, ML) on termi, johon törmää usein myös kyberturvallisuuden alan teosten ja keskustelun ulkopuolella. Näin tapahtuu myös päivittäisissä uutismedioissa ja -artikkeleissa. Tämän pro gradu -tutkielman näkökulmasta on tärkeää määritellä, mitä koneoppimisella tarkoitetaan.

Koneoppiminen voidaan määritellä yhdeksi tekoälyn sekä tietojenkäsittelytieteen osa-alueeksi. Siinä keskitytään datan ja algoritmien käyttöön siten, että tekoäly ja juuri koneoppimismallit kykenisivät ihmisenkaltaiseen oppimistapaan ja täten oppimaan itse parantaen pala palalta tarkkuuttaan. (*What Is Machine Learning (ML)?*, 2024.)

Peruskäsitteenä tietojenkäsittelytieteiden sisällä koneoppiminen sisältää tilastollisen oppimisen sekä optimointimenetelmien käyttämisen. Näiden avulla tietokoneet kykenevät analysoimaan erilaisia tietokokonaisuuksia sekä tunnistamaan kaavoja (engl. patterns). (”What Is Machine Learning (ML)?”, 2020.)

Koneoppiminen voidaan jakaa kahteen eri pääryhmään, jotka ovat valvottu oppiminen (engl. supervised learning) sekä valvomaton oppiminen (engl. unsupervised learning). Valvotulla oppimisella tarkoitetaan tietyn piirrejoukon sekä tavoitearvon välisen suhteen ymmärtämistä. Tavoitearvosta voidaan käyttää myös termejä etiketti (engl. label) tai luokka (engl. class). Konkreettisena esimerkkinä valvottua oppimista voidaan hyödyntää esimerkiksi suhteena, jonka ensimmäisenä osapuolena on luonnollinen henkilö eli ihminen ja hänen demografiset tietonsa. Toisena osapuolena voi toimia ihmisen lainanmaksukyky, ja näiden kahden välistä suhdetta mallinnetaan. Taulukossa 6 kuvataan edellä mainittuja muuttujia. (Saleh & Sen, 2018, s. 33.)

TAULUKKO 6 Henkilön demografisten tietojen ja lainanmaksukyvyn välinen suhde (Saleh & Sen, 2018, s. 33).

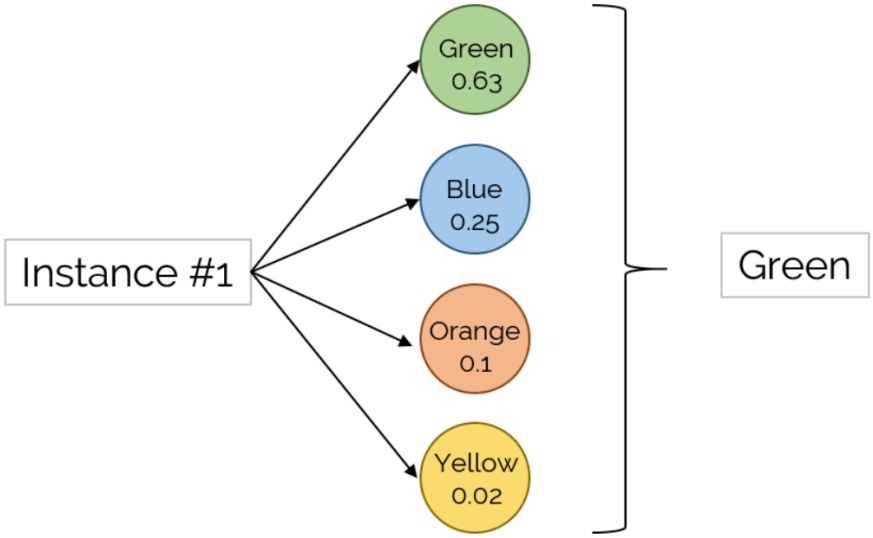
Age	Sex	Education level	Income level	Marital status	Previous loan paid
30	Female	College	$97.000	Single	Yes
53	Male	HIgh school	$80.000	Signle	No
26	Male	Masters	$157.000	Married	Yes
35	Female	None	$55.000	Married	No
44	Female	Undergrad	$122.000	Single	Yes

Tiivistetysti koneoppimista ja sen harjoituttamista voidaan kuvata siten, että taulukon 6 sisältämien muuttujien suhteita ennakoimaan koulutetaan jokin koneoppimismalli. Sen jälkeen tätä samaa koneoppimismallia voidaan soveltaa uuden datan ja tiedon ennustamiseen. Esimerkiksi pankit voisivat hyödyntää tämän kaltaisia malleja ja ne voisivat tällaista apuna käyttäen harkita, myöntävätkö ne lainaa tietylle henkilölle. Samalla ne voivat ennustaa, kuinka suurella todennäköisyydellä tämä henkilö maksaisi lainansa takaisin. (Saleh & Sen, 2018, s. 33.)

Kyberturvallisuuden kontekstissa koneoppimista voidaan hyödyntää esimerkiksi erilaisten petosten tunnistamiseen ja anomalioiden havaitsemiseen. Esimerkiksi pankit ja rahoituslaitokset voivat hyödyntää koneoppimiseen perustuvia tekoälymalleja epäilyttävien tapahtumien havainnointiin. Seuraavassa käsiteltävän valvotun oppimisen avulla voidaan kouluttaa tekoälymalli siten, että sen koulutusdataksi syötetään tietoja tunnetuista petollisista ja haitallisista maksutapahtumista. Kun havaitaan jokin anomalia, kyetään poimimaan epätyypillisiä tapahtumia ja täten voidaan käynnistää lisää tutkimuksia. (*What Is Machine Learning (ML)?*, 2024.)

4.1.4 Valvottu oppiminen

Valvotun oppimisen mallit voidaan jakaa luokittelu- (engl. classificaiton task) - ja regressiotehtäviin (engl. regression task). Käsitellään molemmat tehtävät lyhyesti seuraavaksi. Ensimmäisenä tehdään katsaus luokittelutehtäviin. Luokittelutehtäviä käytetään esimerkiksi jonkin asian, kuten edellä olleen lainan takaisinmaksun ennustamiseen. Tällaiset koneoppimismallit rakennetaan siten, että niille syötetään dataa, joka on luokiteltu erilaisin merkein (engl. label). Luokkia, joita käytetään voi olla enemmän kuin kaksi, mutta niitä on oltava kuitenkin rajattu määrä. Ennustuksen kohteena voisi toimia esimerkinomaisesti formulakuljettajan sijoitus kilpailussa. Kilpailussa ajaa 16 kuljettajaa, joten luokkia olisi täten 16. Luokittelutehtävät pohjautuvat siis ennusteen todennäköisyyteen, joka on havainnollistettu kuviossa 5. (Saleh & Sen, 2018, s. 33–34.)



KUVIO 5 Luokittelualgoritmi yksinkertaistettuna (Saleh & Sen, 2018, s. 34).

Luokittelualgoritmeja on monia erilaisia. Paljon käytettyjä ja muutamia yleisimpiä ovat seuraavat:

- Päätöspuut (engl. decision tree): Kyseinen algoritmi on nimensä mukaisesti puumainen arkkitehtuurinsa osilta. Päätöksentekoprosessin simulointi pohjautuu aina edelliseen päätökseen.
- Naiivi Bayes (engl. naive bayes) -luokittelija: Kyseisenlaiset algoritmit pohjautuvat ryhmään todennäköisyysyhtälöitä, jotka perustuvat Bayesin teoreemaan. Yhtälöt lasketaan olettaen, että eri ominaisuudet (engl. features) eivät ole toisistaan riippuvia. Täten voidaan ottaa huomioon useita ominaisuuksia.
- Neuroverkot (engl. artificial neural network, ANN): Neuroverkot jäljittelevät biologisen neuroverkon rakennetta sekä suorituskykyä ja ne koostuvat toisiinsa liitetyistä neuroneista, jotka ovat sijoitettu ennalta määritellyn arkkitehtuurin mukaisesti. Neuronit välittävät tietoa toisilleen niin kauan, kuin haluttu tulos on saavutettu. (Saleh & Sen, 2018, s. 34.)

Valvottujen mallien toisella laidalla ovat regressiotehtävät (engl. regression task). Niitä käytetään dataan, jonka merkinnät (engl. labels) sisältävät jatkuvia suureita. Tällä tarkoitetaan sitä, että regressiotehtäviä voidaan hyödyntää esimerkiksi joidenkin hyödykkeiden, kuten autojen hintojen ennustamiseen. Täten arvoa ei edusta joukko mahdollisia tuotoksia, vaan jokin suure. Mallin tulosmerkinnät (engl. output labels) voivat olla joko kokonaislukutyyppisiä tai sitten liukulukuja. Väitetään, että kaikista käytetyin regressiotehtävien algoritmi on lineaarinen regressio, joka koostuu ainoastaan yhdestä riippumattomasta piirteestä (x), jonka suhde riippuvaan piirteeseen (y) on lineaarinen. Lineaarinen regressio on yksinkertainen algoritmi. Se toimii hyvin dataongelmissa, jotka ovat yksinkertaisia. Monimutkaisempiin regressioalgoritmeihin lukeutuvat esimerkiksi regressiopuut (engl. regression tree), tukivektoriregressio (engl. support vector regression) sekä neuroverkot (ANN). (Saleh & Sen, 2018, s. 35.)

Valvotun oppimisen malleista voi olla apua esimerkiksi organisaatioille erilaisten reaalimaailman ongelmien ratkaisuissa. Kyberturvallisuuden kontekstissa tällaisia malleja voidaan hyödyntää esimerkiksi roskapostien lajitteluun. (*What Is Machine Learning (ML)?* 2024.)

4.1.5 Valvomaton oppiminen

Toinen koneoppimismallien kategoria on valvomaton oppiminen (engl. unsupervised learning). Valvomattoman oppimisen malli mallinnetaan haluttuun dataan ilman, että sillä on suhdetta tulosmerkintään (engl. output label). Tästä datasta käytetään myös nimeä merkitsemätön data (engl. unlabeled data). Valvomattoman oppimisen alle nivoutuvat algoritmit tähtäävät datan ymmärtämiseen ja siitä mallien löytämiseen. Tätä oppimismallia voidaan hyödyntää esimerkiksi jonkin tietyn asuinalueen ihmisten profiilin ymmärtämiseen. (Saleh & Sen, 2018, s. 35.)

Valvomaton oppiminen voidaan tiivistää siten, että koneoppimisalgoritmeja käytetään kyseisissä tapauksissa analysoimaan ja klusteroimaan merkitsemättömiä datakokonaisuuksia (engl. unlabeled dataset). Algoritmit kykenevät löytämään esimerkiksi piilotettuja kuvioita (engl. pattern) ilman, että ihminen puuttuu asiaan. Tämä on eräs seikka, josta termi valvomaton juontuu. (*What Is Machine Learning (ML)?* 2024.)

Valvomattomaan oppimiseen perustuvat mallit jaotellaan eri tehtäviin samaan tapaan kuin valvotun oppimisen mallit. Suosituin malli on klusterointitehtävä (engl. clustering task), jossa luodaan dataryhmiä eli klustereita. Samaan aikaan edellytetään ehdon noudattamista, jonka mukaan muiden dataryhmien datapisteet (engl. instances) eroavat selkeästi kyseisen ryhmän sisällä olevista datapisteistä. Kaikkien erityyppisten kulsterointialgoritmien tuloksena muodostuu merkintä (engl. label), joka osoittaa instanssin kyseistä merkintää vastaavaan klusteriin. (Saleh & Sen, 2018, s. 36.) Eniten käytettyjä klusterointialgoritmeja ovat:

- K-means, joiden päätehtävä on erottaa datapisteet n-määrään klustereita, joilla on yhtä suuri varianssi keskenään. Tämä tapahtuu minimoimalla kahden pisteen välisten neliöityjen etäisyyksien summa.
- Mean-shift-klusterointi, jossa luodaan klustereita käyttämällä keskipisteitä (engl. centroids). Tässä algoritmimallissa jokainen datapiste on ehdokas keskipisteeksi, joka on kyseessä olevan klusterin pisteiden keskiarvo.
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) on algoritmi, joka määrittää klusterit alueiksi, joilla on suuri pistetiheys. Lisäksi alueet on erotettu alueista, joiden pistetiheys on matala. (Saleh & Sen, 2018, s. 36.)

Saleh & Sen (2018, s. 37) toteavat, että koneoppimisessa on kyse datan ymmärtämisestä. Osa koneoppimismalleista pohjautuu monimutkaisiin matemaattisiin malleihin osan ollessa taas hieman yksinkertaisempia (Saleh & Sen, 2018, s. 37.)

4.1.6 Syväoppiminen

Termistö ja käsitteistö tekoälyn sekä koneoppimisen ympärillä on moninaista ja kirjavaa, eikä ole lainkaan tavatonta, että niistä puhutaan usein päällekkäin. On hyvä huomioida, että samaan aikaan käsitteiden kanssa voi ilmentyä sekaannusta.

Molemmilla, syväoppimisella (engl. deep learning, DL) sekä koneoppimisella on omat nyanssinsa. On tärkeää ymmärtää, että koneoppiminen, syväoppiminen sekä neuroverkot nivoutuvat kaikki tekoälyn osa-alueiksi. On esitetty jaottelu, jonka mukaan neuroverkot ovat koneoppimisen osa-alue ja syväoppiminen on taas neuroverkkojen osa-alue. (*What Is Machine Learning (ML)?*, 2024.)

Suurin eroavaisuus syväoppimisen ja koneoppimisen välillä on siinä, miten kulloinkin käytetty algoritmi oppii. Syväoppimisen prosessissa voidaan ottaa vastaan strukturoimatonta dataa siten, että se on raakamuodossa. Tällainen data voi olla esimerkiksi tekstiä tai kuvia. Syväoppimismalli voi määrittää täysin automaattisesti joukon piirteitä (engl. features), jotka erottavat dataluokat toisistaan. Täten päästään tilanteeseen, jossa voidaan karsia prosessista ihmisen tekemää manuaalista työtä ja voidaan käsitellä tehokkaasti huomattavasti suurempia datamassoja. Syväoppimista on kuvattu skaalautuvana koneoppimisena. Niin kutsuttu ”ei-syvä” koneoppiminen on sen sijaan riippuvaisempi ihmisen toiminnasta, mitä tulee käytetyn mallin oppimiseen. Tärkeä yksityiskohta syväoppimisesta puhuttaessa on lisäksi termin ”syvä” merkitys. Sillä viitataan ainoastaan käytetyn neuroverkon kerrosten lukumäärään ja onkin todettu, että sellainen neuroverkko, joka sisältää yli kolme kerrosta sisältäen tulon (engl. input) ja lähdön (engl. output) voidaan kutsua syväoppimisalgoritmiksi tai syväksi neuroverkoksi. (*What Is Machine Learning (ML)?*, 2024.)

4.2 Tekoäly kyberhyökkäyksissä

Tässä luvussa käsitellään tekoälyn hyödyntämistä osana hyökkäyksellisiä ja vihamielisiä toimia kybertoimintaympäristössä. Aluksi pureudutaan siihen, millaisia hyötyjä tekoäly tässä kontekstissa tuo. Sitä kautta siirrytään tekoälyavusteisten kyberhyökkäysten piirteisiin ja vaiheisiin. Loppupuolella käsitellään, millaisia eri hyökkäystyyppejä on, ja miten eri uhkatoimijat käyttävät tekoälyä. Viimeisenä tehdään pieni katsaus ennusteeseen tekoälyn roolista lähitulevaisuudessa.

Ventren mukaan (2020, s. 158–159) kyberhyökkäyksien rooli on hyvin merkittävä, kun tarkastellaan strategisia haasteita. Hyökkäysvektorit ovat moninaisia ja tapoja sekä menetelmiä toteuttaa vihamielisiä toimia kyberavaruudessa on paljon. Tekoälyllä on potentiaalia ja sitä voidaan hyödyntää sekä kyberhyökkäyksissä että samaan aikaan puolustuksellisissa toimissa. Täten asetelma on ajautunut kaksintaisteluun ja vastakkainasetteluun, kuten myös yhtäältä voimatasapainon keikkumiseen hyvän ja pahan välillä. Oli toimija sitten lain ja moraalin oikealla tai väärällä puolella, kysymyksiä siitä, mikä tekoälyn

rooli kunkin pelaajan työkalupakissa on, esitetään suurella todennäköisyydellä. (Ventre, 2020, s. 158–159.)

Ventre (2020, s. 159) toteaa, ettei tekoäly itsessään ole tärkein asia. Hänen mukaansa tekoälyn syvin olemus on, mitä se tarjoaa kullekin käyttäjälle. Samaan aikaan tekoäly edustaa monimutkaisuutta sekä hyökkääjän että puolustavan osapuolen näkökulmasta. Ventre (2020, s. 159) sanookin, että monimutkaisuus ja kompleksisuus voidaan nähdä synonyyminä laadulle ja tehokkuudelle. Vaikka monimutkaisuus lisääntyisikin, ei se poista tosiasiaa, että kaksintaistelu ja etumatkan saavuttaminen molempien osapuolten välillä säilyy. Samalla on tärkeää ymmärtää ja tiedostaa, että monimutkaisuus kuuluu ja on osa kybertoimintaympäristöä ja yhtäältä alan toimijat ovat tottuneet elämään sen kanssa. Se ei siis ole lainkaan uusi muuttuja, vaikka tekoäly on vallannutkin alaa ja vauhti tulee vain kiihtymään. (Ventre, 2020, s. 159.)

Tekoälyllä on varjopuolensa, kuten monilla ihmiskuntaa ja yksittäisiä ihmisiä hyödyttäneillä asioilla ja esineillä on. Sitä voidaankin käyttää vihamielisiin ja toista ihmistä tai organisaatiota vahingoittaviin tarkoituksiin. Kuten Ventre (2020, s. 162) esittää, mikä tahansa työkalu, jonka käyttö poikkeaa alkuperäisestä ja pääasiallisesta käyttötarkoituksestaan, voidaan muuttaa vahingoittamis- ja tuhoamisvälineeksi. Esimerkiksi puun pala voidaan muuttaa aseeksi, jos sillä lyödään toista ihmistä. Täten se muuttuu aseeksi nimenomaan käytön kautta. (Ventre, 2020, s. 162.)

Tekoälyn avulla voidaan toteuttaa lukuisia erilaisia sovelluksia ja tehtäviä. Sillä onkin moninaisia kykyjä kuten päättely- ja ongelmanratkaisukykyä. Se voi myös keksiä tarkoituksia ja merkityksiä sekä yleistää ja oppia erilaisista kokemuksista, joita sille syötetään. Ikävä puoli on, että vihamieliset toimijat käyttävät näitä älykkäitä kykyjä, kun he suunnittelevat ja toimeenpanevat hyökkäyksellisiä kybertoimia kohdeorganisaation tai -henkilön tietojärjestelmiä vastaan. Täten tekoäly voidaan valjastaa aseeksi.

Myös Guembe ym. (2022) näkevät tekoälyn voimakkaan tulemisen merkittävänä ilmiönä osana uhkatoimijoiden työkalupakkia kyberhyökkäyksissä. On käytetty jopa sellaista käsitettä kuin ”hyökkäävä tekoäly”. Sillä tarkoitetaan uhkatoimijoiden suunnattuja ja kohdennettuja hyökkäyksiä, joiden nopeus sekä mittakaava on hyvin suuri. Samaan aikaan uhkatoimijat kykenevät hyökkäävän tekoälyn ansiosta välttämään tai kiertämään perinteiset sääntöihin pohjautuvat havaitsemistoimenpiteet. Ilkeämielisten toimijoiden käsissä tekoälyyn liittyy mittava potentiaali sen ansiosta, että sillä on hedelmälliset mahdollisuudet oppia ja sopeutua kohdeympäristössään. Tämä taas mahdollistaa skaalautuvat, räätälöidyt ja ihmisenkaltaiset hyökkäyksen ja niiden myötä aivan uuden aikakauden. On myös esitetty väite, jonka mukaan nykyiset puolustukselliset kyberturvallisuusratkaisut eivät enää tehoa kehittyneitä tekoälyavusteisia kyberaseita vastaan. (Guembe ym., 2022, s. 2377.)

4.2.1 Tekoälyn luomat hyödyt kyberhyökkäyksissä

Voidaan nähdä, että tekoäly on uusi ja nouseva trendi osana kyberhyökkäyksiä. Chakraborty ym. (2022, s. 15) mainitsevatkin, että ei voida vielä kuitenkaan

varmuudella todeta, mitkä sen todelliset vaikutukset ovat verkkorikollisuuden tulevaisuuteen. Huomioitavaa on, että tekoälyn rooli on hyvin moninainen, kun tarkastellaan offensiivisia kykyjä sekä strategisesta että taktisesta näkökulmasta (Sharikov, 2018).

Se miten tekoäly hyödyttää kyberhyökkäyksiä, voidaan hahmottaa suhteellisen helposti. Kaikessa yksinkertaisuudessaan tekoälyn avulla voidaan hyökkäävissä toimenpiteissä automatisoida sellaisia tehtäviä, jotka jouduttaisiin muuten tekemään manuaalisesti ihmisen toimesta. Tällaisia voivat olla esimerkiksi haavoittuvuuksien etsiminen sekä löytäminen ja sittemmin niiden hyödyntäminen kyberhyökkäyksessä. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 7)

Edellä kuvattu yksinkertaistus ei ole kuitenkaan koko totuus, mutta tällaisen kuvauksen avulla saadaan kattavahko yleiskäsitys, mistä on kyse. On hyvä huomioida, että tekoälysovellusten käyttöönotto on ollut hidasta kyberturvallisuuden ratkaisuissa, jolloin samaan aikaan hyökkääjät ovat kyenneet olemaan tuottavia ja edistyneet omissa toimissaan tekoälyn osalta (Turtiainen ym., 2023, s. 126).

Kyberhyökkäyksien suunnittelu ja toteutus vaatii paljon vaivaa ja ammattitaitoa sekä lisäksi erilaisia työkaluja. Se, miten tekoäly auttaa uhkatoimijoita toimissaan voidaan jakaa karkeasti kolmeen seikkaan. Ensimmäiseksi tekoäly tuo nopeutta. Kuten aiemmin mainittua, tekoälyn avulla voidaan korvata sellaisia työvaiheita, jotka on aiemmin jouduttu tekemään käsin. Tämä luonnollisesti nopeuttaa toimintaa huomattavasti, kuten muissakin sovelluskohteissa laillisin tarkoitusperin tehdyissä toimissa. Hyökkäyksen kannalta vaadittavia toimia on esimerkiksi tunnistetietojen hankkiminen, haavoittuvuuksien etsiminen sekä salasanojen murtaminen. Iso osa tällaisista toimista voidaan tekoälyn avulla automatisoida. Näin ollen hyökkääjä säästää aikaa eikä sen tarvitse olla kohdejärjestelmässä niin kauaa, jonka seurauksena riski kiinnijäämisestä muuttuu pienemmäksi. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 9.)

Toinen etu on tehokkuus, jonka tekoäly mahdollistaa. Hyökkäyksiä voidaan skaalata entistä suuremmiksi ja näin ollen ne voivat aiheuttaa vakavampia seurauksia. Yhtäältä automatisoituja hyökkäyksiä on mahdollista käynnistää samanaikaisesti määrällisesti useampia sekä useampaa kohdetta kohtaan. Kaikki tämä voidaan myös tehdä lyhyemmän aikaikkunan sisällä. Konkreettisena esimerkkinä voitaisiin käyttää erityisen räätälöityä hyökkäystä, johon lukeutuu muun muassa kohdennettu tietojenkalastelu. Myös hyökkäysten tarkkuutta saadaan korkeammaksi tekoälyn avulla. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 9.)

Turtiainen ym. (2023, s. 126) näkevät skaalautuvuuden mahdollisuudet rajattomina tekoälyteknologioiden osalta. Samaan aikaan hyökkäävää osapuolta edesauttava seikka on tekoälypohjaisten hyökkäysten suoritusnopeus ja sen merkittävä kasvu. Se asettaa puolustavalle osapuolelle lisähaastetta reagoida tällaisiin hyökkäyksiin, koska reagointiaikaa on luonnollisesti vähemmän. (Turtiainen ym., 2023, s. 126.)

Kolmanneksi tekoälyn voimin hyökkäyksien kattavuutta saadaan parannettua, jolloin ne ovat myös laaja-alaisempia ja kokonaisvaltaisempia. Esimerkiksi tekoälyä voidaan hyödyntää kohteen tiedustelussa analysoimalla suuria

määriä avoimista lähteistä löytyvää dataa. Täten voidaan löytää uusia hyökkäyspolkuja ja saavuttaa pääsy isompaan osaan tavoitelluista kohteista. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 9.)

Myös Turtiainen ym. (2023, s. 126) ovat tekoälyn tuottamien etujen osalta samoilla linjoilla. He mainitsevat tekoälyn ratkaisevina etuina sekä eettisille että pahantahtoisille hakkereille asioita, kuten kyvykkäiden toimijoiden määrän lisääntyminen, mahdollisten hyökkäysten tiheyden laajamittainen kasvu sekä mahdollisten hyökkäyskohteiden moninaisuuden lisääntyminen (Turtiainen ym., 2023, s. 126).

Turtiainen ym. (2023, s. 126) toteavatkin, että edellisessä kappaleessa mainitut saavutukset ovat mahdollisia siitä syystä, että tekoälyteknologioihin pohjautuvat järjestelmät ovat skaalautuvia, kustannustehokkaita sekä niiden käyttöönoton helppouden vuoksi. He (Turtiainen ym., 2023, s. 126) mainitsevat datamäärien massiivisen kasvun tulevissa kyberavaruudessa käytävissä ”sodissa” sekä sen, kuinka tekoälyteknologiat hyödyttävät hyökkääjiä tiedustelussa, kohteeseen tunkeutumisessa, tietojen poistamisessa sekä käyttöoikeuksien laajentamisen vaiheissa. Uhkatoimijoiden käyttämänä ja hyödyntämänä tekoäly muodostaakin merkittäviä uhkakuvia puolustaville osapuolille ja organisaatioille. (Turtiainen ym., 2023, s. 126.)

4.2.2 Tekoälyavusteisten kyberhyökkäysten piirteet

Tekoälyn avulla hyökkääjillä on mahdollisuus olla toimissaan entistä hienovaraisempia, koska tekoäly avaa ovia kehittyneempiin työkaluihin ja edelleen älykkäämpään automaatioon. Kehittyneisyys ilmenee kolmella tapaa (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 10).

Ensimmäisenä ilmentymänä esiintyy kontekstualisointi. Tekoälyllä voidaan oppia tehokkaasti kohdejärjestelmästä ja hyödyntää sieltä saatua tietoa vaikkapa haittaohjelman luomiseen, jota sittemmin käytetään kyseistä järjestelmää vastaan. Hyökkäyksen myöhemmissä vaiheissa on mahdollista esimerkiksi räätälöidä haitallinen tietoliikenne toimimaan huomaamattomammin kohdeverkossa, tai haittaohjelman käytös kyetään optimoimaan siten, että se ikään kuin sekoittuu normaaliin toimintaan kohdejärjestelmän sisällä. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 10.)

Toinen ilmentymä on muovautuvuus, jolla tarkoitetaan tässä tapauksessa tekoälyn kykyä oppia ja uudelleenoppia kohdeympäristö automaattisesti. Voidaan siis nähdä, että hyökkäykset ovat osiltaan autonomisia, sillä ne voivat muokkautua ympäristöönsä ja siellä tapahtuviin muutoksiin itsenäisesti. On hyvä huomioida, että muovautuvuudella tarkoitetaan ennemminkin pysyvää kuin kertakäyttöistä ominaisuuttaa kyberhyökkäyksissä, joissa hyödynnetään tekoälyä. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 10.)

Guemben ym. (2022) näkemys myötäilee tekoälyn mahdollistamaa muovautuvuutta ja heidän arvionsa mukaan tulevaisuudessa hyökkäysmenetelmät voivat olla hyvinkin tietoisia toimintaympäristöstään. Täten kyseisen kaltaiset menetelmät pystyvät tekemään itsenäisesti päätöksiä kohdeympäristönsä perusteella. Kuten luvussa 3.3.2 mainittiin, APT-ryhmät pyrkivät säilyttämään

jalansijansa mahdollisimman pitkään kohdejärjestelmässä, johon kyberhyökkäys kohdistuu. Onkin APT-toimijoiden osalta huomionarvoista tekoälyn mahdollistaman sopeutumisen myötä, että mitä pitempään tekoälyn avulla toteutettu hyökkäys pääsee olemaan kohdejärjestelmässä ja isännässä (engl. host), sitä tehokkaammaksi integroituminen kohteeseen muuttuu. Lisäksi pitkään kohdejärjestelmässä pysyminen mahdollistaa riippumattomuuden ympäristöstä sekä kyberturvallisuuden puolustusinfrastruktuurin toimista hyökkääjää kohtaan. (Guembe ym., 2022, s. 2377)

APT-ryhmät ovat kykyjensä ja resurssiensa puolesta yksi vakavimmista kyberuhkien aiheuttajista valtiollisille sekä yksityisille organisaatioille, ja tätä eivät puolustavan osapuolen näkökulmasta helpota edellä mainitut seikat lainkaan. Jo itsessään kehittyvien tekoälyyn perustuvien hyökkäystekniikoiden seuraukset voivat pahimmillaan olla erittäin tuhoisia ja jopa hengenvaarallisia (Guembe ym., 2022, s. 2377).

Kolmantena tekoälyavusteiset kyberhyökkäykset voivat olla hyvin vaikeasti havaittavissa verrattuna perinteisiin hyökkäyksiin. Tällä on vaikutus hyökkäyksen resilienssin kasvuun positiivisella tavalla. Tekoälyn avulla voidaan älykkäästi ja optimoidusti toteuttaa tiedusteluun liittyvää data-analyysia sekä tiedonhakua. Yksi uhkatoimijan kiinnijäämiseen liittyvä riski on sen käyttämän kohdejärjestelmään asennetun haittaohjelman ja uhkatoimijan hallinnoiman komentopalvelimen välinen kommunikaatio ja tietoliikenne. Tekoäly mahdollistaa haittaohjelmien korkeamman itseohjautuvuuden, jonka johdosta sen ei tarvitse kommunikoida niin aktiivisesti komentopalvelimen kanssa. Näin ollen riski vihamielisen ja haitallisen toiminnan havaitsemiseksi laskee. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 10.)

4.2.3 Kyberhyökkäyksen vaiheet ja tekoäly

Erityisesti APT-hyökkäykset ovat usein hienostuneita sekä monimuotoisia, ja niiden kuvaamista varten on ollut selkeä tarve kehittää erilaisia viitekehysmalleja. Mikäli hyökkäävällä osapuolella, eli uhkatoimijalla, on runsaasti taitoa ja päättäväisyyttä, se voi käyttää hyökkäystä tehdessään useita hyökkäysvektoreita sekä sisääntulopisteitä navigoidakseen puolustuksen ympärillä. Hyvin resursoitu ja taidokas, vihamielinen osapuoli voi toteuttaa hyökkäyksen niin hyvin, että toimia ei huomata kuukausiin tai jopa vuosiin. (Lehto, 2022, s. 122.)

Kyberhyökkäyksien eri vaiheita varten on luotu erilaisia viitekehyksiä ja malleja, jotka helpottavat niiden tutkimista ja havainnollistamista. Kyberhyökkäysten prosessi voidaan havainnollistaa elinkaarena. Erilaisia APT-hyökkäyksiä on nykyään hyvin laaja skaala, joten on nähty tarve kuvata hyökkäysprosesseja erilaisten viitekehysten avulla. Yleisimpiä viitekehyksiä ovat:

- MITRE ATT&CK,
- Mandiant Attack Life Cycle Model,
- Lockheed Martin Cyber Kill Chain,
- Unified Kill Chain,
- Hybrid Kill Chain. (Lehto, 2022, s. 122.)

Yksi alalla paljon käytetty viitekehys on yhdysvaltalaisen aseteollisuuskonserni Lockheed Martinin luoma Cyber Kill Chain. Se on muotoutunut asevoimissa käytetystä tappoketjusta ja se on lähestymistapa, jossa lähestytään vaiheittain vihollisen tunnistamista ja pysäyttämistä. Tappoketju on tarkoitettu puolustautumiseen pääasiassa APT-toimijoita ja niiden tekemiä kehittyneitä hyökkäyksiä vastaan. APT-toimijat käyttävät paljon ajallisia resursseja hyökkäyksen suunnitteluun sekä tarkkailutoimintaan. (*What Is the Cyber Kill Chain?*, ei pvm.)

Lockheed Martinin tappoketjumalli on seitsenvaiheinen ja sen vaiheet ovat: tiedustelu (engl. reconnaissance), aseistus (engl. weaponization), toimitus (engl. delivery), hyväksikäyttö (engl. exploitation), asennus (engl. installation), komento ja hallinta (engl. command & control) ja tavoitteisiin tähtäävät toimet (engl. actions on objectives). (*What Is the Cyber Kill Chain?*, ei pvm.) Kuviossa 6 (*Cyber Kill Chain®*, ei pvm.) on havainnolistettu Cyber Kill Chain -viitekehys.

KUVIO 6 Cyber Kill Chain (*Cyber Kill Chain*®, ei pvm.)

Mandiant Attack Life Cycle Model -viitekehys kuvaa APT-hyökkäysten syklistä toimintamallia. Kyseisessä viitekehyksessä hyökkäykset on jaoteltu kahdeksaan eri vaiheeseen. Unified Kill Chain -viitekehys on puolestaan rakennettu siten, että siinä on yhdistelty elementtejä Cyber Kill Chain ja MITRE ATT&CK-viitekehyksistä. Unified Kill Chain koostuu 18 eri vaiheesta ja sen päällimmäisenä tavoitteena on laajentaa Cyber Kill Chain -mallia. Yhtäältä sen on tarkoitus olla parannus MITRE ATT&CK:n aika-agnostiseen luonteeseen verrattuna. Mallin fokus on sellaisissa taktiikoissa, jotka muodostavat kyberhyökkäyksen perättäiset vaiheet. (Lehto, 2022, s. 123–124.)

Paljon käytetty viitekehys on edelläkin esiin noussut MITRE ATT&CK, jossa kyberhyökkäyksen eri vaiheet ja hyökkääjän tavoitteet on jaoteltu 14 eri vaiheeseen. Näihin vaiheisiin sisältyvät kokonaisvaltaisesti tekniikat ja taktiikat, joita kyberhyökkäyksissä käytetään. Näitä ovat kyseisen viitekehyksen mukaan esimerkiksi tiedustelu (engl. reconnaissance), ensimmäinen pääsy (engl. initial access), läsnäolon ylläpito (engl. persistence), puolustuksen välttely (engl. defense evasion), tunnuksien saanti (engl. credential access), tunkeutumisen laajentaminen (engl. lateral movement) ja niin edelleen. (*MITRE ATT&CK®*, ei pvm.)

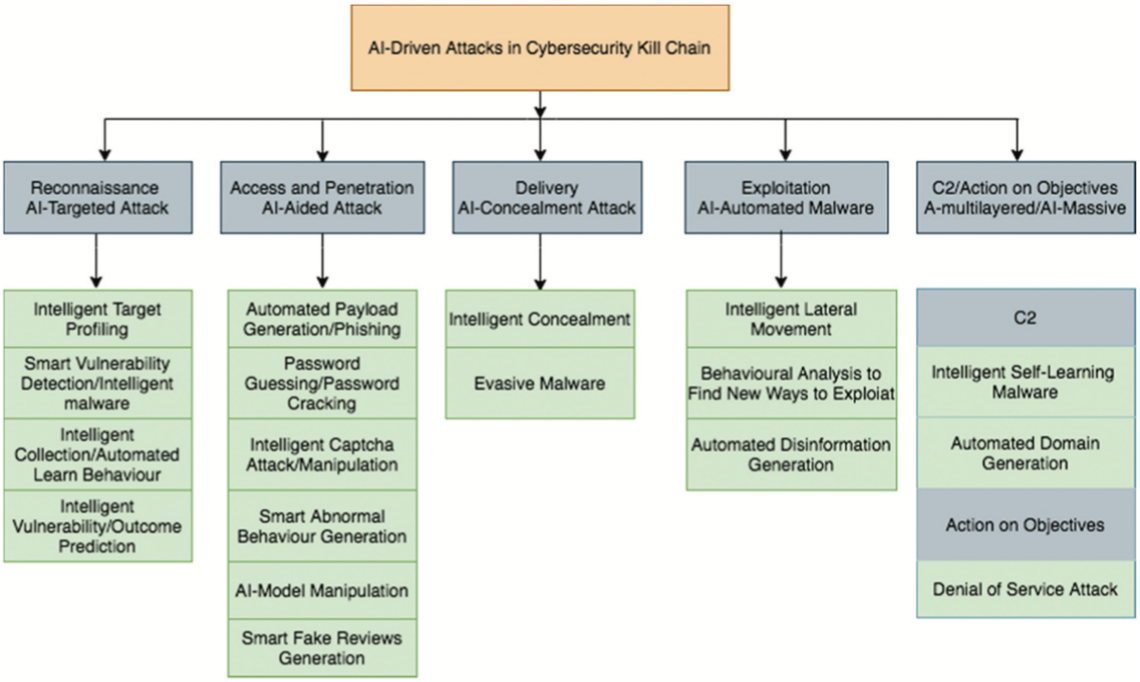
Hyökkääjät hyötyvät tekoälyn käytöstä useissa edellä mainituissa vaiheissa ja samanaikaisesti uhkatoimijat voivat tekoälyä käyttäen luoda uusia tekniikoita tavoitteidensa saavuttamiseksi. Kyvykkyydet, joita tekoäly kyberhyökkäyksissä mahdollistaa, voidaan jaotella kuuteen eri luokkaan: (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 13.)

- Automaatio,
- Hyökkäyksen vaikeasti havaittavuus,
- Hyökkäyksen jatkuvuus,
- Käyttäjän manipulointi,
- Käyttäjätunnusten varastaminen,
- Tiedonkeruu.

Tekoälyä voidaan hyödyntää useissa eri kyberhyökkäyksen vaiheissa. Myös Turtiainen ym. (2023, s. 127) mainitsevat, että tekoäly ja erityisesti koneoppimisteknologiat ovat ottaneet paikkansa kyberhyökkäyksissä useissa eri tappoketjun vaiheissa. Heidän mukaansa hyökkääjät käyttävät tekoälyä pääasiallisesti tiedustelussa, tunkeutumisessa, sivuttaisliikkeessä järjestelmän sisällä sekä tavoitteisiin tähtäävissä toimissa. Tekoälyn kehittyminen kuitenkin mahdollistaa sen käytön koko ajan enenevissä määrin myös hyökkäyksen muissa vaiheissa. (Turtiainen ym., 2023, s. 127.)

Hyökkäysketjun alkupään vaiheissa uhkatoimijat voivat hyödyntää tekoälyä esimerkiksi kohteen tiedustelun parantamiseen ja kehittämiseen. Tämä mahdollistaa normaalin käyttäytymisen ja toiminnan tutkimisen, joka kohdistuu hyökättäväksi valitun kohteen kyberpuolustusmekanismeihin, tietokoneinfrastruktuureihin sekä laitteisiin. Tekoälyn avulla uhkatoimijalla on paremmat mahdollisuudet saavuttaa rakenteellisia, topologisia ja toiminnallisia tietoja laitteista, verkkoinfrastruktuurista ja verkkoliikenteestä. Näin ollen uhkatoimija kykenee tunnistamaan tehokkaammin kriittiset pisteet kohdejärjestelmästä, johon aiotaan hyökätä. (Guembe ym., 2022, s. 2395.)

Guembe ym. (2022, s. 2383) esittävät havainnollistavan kuvion 7 avulla Cyber Kill Chain -viitekehystä vasten, kuinka heidän tutkimuksessaan tekoälyn käyttötavat ilmenivät kyberhyökkäyksissä. Tappoketjun ensimmäisissä vaiheissa tekoälyä hyödynnetään enemmän kuin loppupään vaiheissa.



KUVIO 7 Hyökkäyksellisen tekoälyn käyttötavat kuudessa kyberturvallisuustappoketjun vaiheessa (Guembe ym., 2022, s. 2383).

Taulukossa 7 on lisäksi eritelty tekoälyn mahdollistamia tekniikoita hyökkäyksen eri vaiheissa Traficomin raportin (2022, s. 17) mukaan.

TAULUKKO 7 Tekoälyn mahdollistamat hyökkäystekniikat (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 17).

Kyberhyökkäyksen vaihe	Tekoälyn mahdollistamat hyökkäystekniikat
Tiedustelu	• Captcha breakerin käyttö • Automatisoidun asiointibotin käyttö
Tunkeutuminen	• SNAP_R avulla kohdennettu tietojenkalastelu • Mechanical phish avulla haavoittuvuuksien skannaaminen
Komentopalvelimen perustaminen	• Viestinnän analysointi • Empirellä verkon liikenteen muovaaminen
Käyttöoikeuksien laajentaminen	• CeWL salasanageneraattori
Tunkeutumisen laajentaminen	• Automatisoitu tekoälyn mahdollistama hyökkäyksen suunnittelu • Automatisoitu toteutus MITRE CALDERAn avulla
Tiedon vieminen / varastaminen	• Arvokkaan tiedon tunnistaminen automaattisesti • Viestinnän analysointi • Empirellä verkon liikenteen muovaaminen

4.2.4 Tekoälyn mahdollistamat hyökkäyskyvyt

Hyökkäyskyvyt, joita tekoäly mahdollistaa, ovat moninaiset. Samaan aikaan on paljon mystisen verhon peitossa. On olemassa kalliita, fyysisesti suuria ja suorituskykyisiä supertietokoneita, joilla on mittavat laskentaresurssit, jota tekoälyn suuren mittakaavan hyödyntäminen edellyttää. Tällaisten koneiden avulla voidaan hyökätä esimerkiksi strategisesti tärkeisiin sotilaskohteisiin tai kriittisen infrastruktuurin kannalta tärkeitä organisaatioita vastaan. Samalla pyritään häiritsemään niiden toimintaa. Tiedetään, että tällaisia hyökkäyksellisiä kykyjä on, mutta niiden tutkiminen on hyvin haastavaa niihin liittyvän tiedon salaisuuden vuoksi. (Sharikov, 2018, s. 370.)

Taulukossa 8 on jaoteltu edellisessä alaluvussa luetellut kuusi tekoälyn mahdollistamaa hyökkäyskykyä. Taulukosta voidaan todeta, että tällaiset kyvyt tarjoavat hyötyä eniten kyberhyökkäysten tiedustelu- ja puolustuksenvälttelyvaiheita. Traficomin (2022, s. 13) raportin mukaan tällä hetkellä käytössä olevat tekoälytekniikat eivät tarjoa vielä merkittäviä hyötyjä käyttöoikeuksien korottamiseen, suorittamiseen tai läsnäolon ylläpitoon. Tekoälyn hyökkäysominaisuudet antavat siis eniten hyötyjä hyökkäysketjun ensimmäisiin ja loppupään vaiheisiin. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 13.)

TAULUKKO 8 Tekoälyn mahdollistamat hyökkäyskyvyt (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 13).

Tekoälyn mahdollistama kyvykkyys	Kyberhyökkäystaktiikka	Tekoälyn mahdollistama hyökkäystekniikka
Automaatio	• Tiedustelu • Ensimmäinen pääsy • Tunkeutumisen laajentaminen • Vaikuttaminen	• Hyökkäysten muovautuminen • Hyökkäysten koordinointi • Hyökkäyskampanjat • Haavoittuvuuksien löytäminen
Vaikeasti havaittavuus	• Tiedustelu • Ensimmäinen pääsy • Läsnäolon ylläpito • Tunkeutumisen laajentaminen • Tiedonkeruu • Komentopalvelimen perustaminen • Tiedon vieminen / varastaminen	• Kiinni jäämisen välttäminen • Skannaus • Eteneminen • Tiedon vieminen / varastaminen
Hyökkäyksen jatkuvuus	• Läsnäolon ylläpito • Puolustuksen välttely	• Hyökkäyksen suunnittelu • Haittaohjelman sovittaminen kohteen normaaliin toimintaan • Virtualisoinnin tunnistaminen
Käyttäjän manipulointi	• Tiedustelu • Ensimmäinen pääsy • Käyttöoikeuksien laajentaminen • Tunnusten saaminen	• Kohteen valinta • Kohteen seuranta • Kohdennettu tietojen kalastelu • Imitaatio • Valheellisten profiilien rakentaminen
Käyttäjätunnisteiden varastaminen	• Ensimmäinen pääsy • Tunnusten saaminen	• Biometristen tunnistamis-keinojen ohittaminen • Näppäinpainallusten tunnistaminen • Salasanojen arvaaminen
Tiedonkeruu	• Tiedustelu • Tunnusten saaminen • Tiedonkeruu • Vaikuttaminen	• Tiedon louhinta avoimista lähteistä • Kohteen valinta • Kohteen seuranta • Tiedustelu

Erilaisten kyberhyökkäysten kirjo on nykyään hyvin laaja ja tekoälyä sovelletaan paljon monenlaisiin vihamielisiin toimenpiteisiin. On havaittu, että tekoäly kykenee kasvattamaan hyökkäysten menestymistä. Samaan aikaan tekoälytekniikoiden kypsyys on sellaisella tasolla, että niiden tuoma hyöty voidaan aidosti nähdä. Tämä ilmenee esimerkiksi käyttäjien manipuloinnissa sekä hyökkäyksellisten toimenpiteiden havaitsemisen hankaloittamisessa. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 14.)

4.2.5 Hyökkäystyypit

Hyökkäystyyppejä, joissa tekoälyä käytetään, on useita. Tässä alaluvussa esitellään kaksi esimerkkiä, joista löytyy näyttöä. Tietojenkalastelu on varmasti monelle ihan tavallisellekin internetin käyttäjälle kyberhyökkäystyyppi, joka on

tuttu, ja niistä varoitellaan säännöllisesti medioissa ja työpaikoilla. Niiden uhriksi joutuminen ei olekaan tänä päivänä lainkaan tavatonta, vaikka ei olisikaan mielestään korkean profiilin kohde tai vaikutusvaltaisessa asemassa.

Kohdennettu tietojenkalastelu on yksi yleinen hyökkäyksen tapa ja tekoälyllä voi olla sen toteuttamisessa helpottava rooli. Sen avulla voidaan tukea uhrien valitsemista ja valikoida kohteeksi sellaisia ihmisiä, joilla on uhkatoimijaa kiinnostava ominaisuus, kuten korkea asema kohdeorganisaatiossa. Tällaista toimintaa ja maalittamista kutsutaan termillä käyttäjäprofilointi. Tekoälyn avulla kerätään tietoa julkisista lähteistä, joihin lukeutuvat esimerkiksi sosiaalisen median alustat, kuten Facebook ja LinkedIn. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 14.)

Tekoälyavusteisesti tiedonkeruu voidaan kohdistaa tietyn kohdeorganisaation henkilöstöön. Käyttäjäprofiileista uhkatoimijat haluavat kerätä tietoa seuraajista, ystäväverkostosta, yhteystiedoista, tilin iästä, julkaisujen määrästä, tykkäysten määrästä, uudelleenjulkaisuista, reaktioista, kiinnostusten kohteista sekä vaikkapa harrastuksista. Kun saadaan riittävä määrä tällaista dataa, voidaan luokitella uhreja samankaltaisten käyttäjien kanssa samoihin ryhmiin. Lopuksi käyttäjät pyritään tunnistamaan. Sen jälkeen profiilit käydään läpi luonnollisen kielen käsittelymenetelmiä hyödyntäen (engl. Natural Language Processing, NLP). Prosessoinnin tuloksena saadaan käyttäjää kiinnostavia aiheita, jotka syötetään koulutettuun integroimismalliin. Siten voidaan tuottaa kohdennettua tietojenkalastelua varten käyttäjälle räätälöityjä sähköposteja. Yksi automaattinen tietojenkalastelumalli on tieteen kentällä tutkijoiden suunnittelusta lähtenyt SNAP_R. On kyetty osoittamaan, että SNAP_R:n tekemät twiitit keräävät enemmän klikkauksia kuin ihmisten tekemät. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 14.)

Toinen esimerkki on haittaohjelman toiminnan piilottaminen. Tänä päivänä yksi yleisimmistä kyberhyökkäyksen havaitsemis- ja pysäyttämiskeinoista on havaita haitallinen tietoliikenne tietoverkon sisällä. Uhkatoimijat voivat tekoälyn avulla ujuttaa haitallista ja heidän kannaltaan edullista liikennettä niin sanotun normaalin ja ei-pahantahtoisen liikenteen sekaan. Yleensä kun haittaohjelma on saatu istutettua kohdejärjestelmään, se pysyy passiivisena ja tekee ainoastaan verkkoliikenteen valvontaa kohteessaan. Hyökkääjä voi toimia esimerkiksi siten, että se analysoi verkkoliikennettä. Kun liikenne on ruuhkaisimmillaan, se voi ajoittaa kommunikaationsa komentopalvelimelleen näihin ajankohtiin. Verkkoliikenteen analysoinnilla tekoälyä hyödyntäen uhkatoimijat ovat siis kuin kaloja vedessä oman ilkeämielisen liikenteensä kanssa kohdejärjestelemän normaalin liikenteen seassa. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 16.) Pankkiakaan ei kannata mennä ryöstämään keltaisella avoautolla, vaan ehkä kohdemaassa eniten myydyllä automallilla, joka on maalattu suosituimmalla värillä.

Lisäksi huomionarvoisia hyökkäystyyppejä ja -teknologioita ovat Sharikovin (2018, s. 370) esiin nostamat autonomiset agentit, jotka mahdollistavat tunkeutumisen ja soluttautumisen kohdejärjestelmään. Ne kykenevät muistamaan hyökkäyksen tiedusteluvaiheen aikana kerätyt tiedot ja myöhemmin ne osaavat hyödyntää tätä informaatiota tunkeutumisreitin suunnitteluun. Autonomisista agenteista on mahdollista muodostaa myös parvia, jotka koostuvat

yhteistyössä toimivista autonomisista agenteista. Täten ne voivat muodostaa bottiverkon, joka ei vaadi keskitettyä ohjausta eikä valvontaa. (Sharikov, 2018, s. 370.)

4.2.6 Uhkatoimijoiden erot

Kyberhyökkäyksissä tekoälyä hyödyntävät monenlaiset eri toimijat. Uhkatoimijoilla on hyvin kirjava joukko erilaisia motiiveja hyökkäystensä toteuttamiseksi ja samaan aikaan myös resurssien määrä vaihtelee suuresti. Siihen, miten jokin uhkatoimija käyttää tekoälyä, vaikuttavat toimijan teknologiset valmiudet, tekoälyyn liittyvien taitojen saatavuus sekä luonnollisesti toimijan kiinnostus tekoälyä kohtaan osana sen hyökkääviä toimia. Jotta voidaan ennustaa, milloin ja miten tekoälyä käytetään osana hyökkäyksiä, on ensisijaisen tärkeää ymmärtää uhkatoimijoiden motiiveja ja taustoja. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 19.)

Kyberhyökkäyksissä tekoälyä hyödyntävät uhkatoimijat on jaoteltu kolmeen. Ensimmäisenä ovat yksittäiset hyökkääjät, jotka voivat tuoda tekoälyn osaksi omia toimintamallejaan siten, että sen avulla nopeutetaan ja laajennetaan toimintaa. Yksittäiset tekijät voivat korvata manuaalista työtä automatisoimalla tehtäviä. Samaan aikaan tällaiset hyökkääjät käyttävät enimmäkseen valmiita työkaluja eivätkä niinkään kehitä niitä itse. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 19–20.)

Toisena tekoälyä käyttävänä uhkatoimijasegmenttinä tulevat rikollisryhmät, joiden toiminnassa tekoäly voi näyttäytyä liiketoimintaa optimoivana ja taloudellisia voittoja maksimoivana tekijänä. Rikolliset toimijat voivat hyödyntää tekoälyä ensisijaisesti hyökkäyksien ensimmäisissä vaiheessa. Tästä esimerkkinä toimii arvokkaiden kohteiden identifioiminen. Voidaan todeta, että toiminta, joka nivoutuu kohteen tiedusteluun, voisi olla rikollisia hyödyttävä ja kiinnostava aspekti tekoälyn hyödyntämisessä. Tällaiset uhkatoimijat käyttänevät valmiita työkaluja, mutta heillä voi olla myös kykyjä ja resursseja kustomoida niitä. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 19–20.)

Kolmantena ovat valtiolliset uhkatoimijat. Tällaisten toimijoiden nähdään hyödyntävän kehittyneempiä tekoälyteknologioita. Valtiolliset uhkatoimijat voivat hyödyntää niitä suorina ja piilotettuina mekanismeina. On todettu, että kyseiset toimijat voivat käyttää tekoälyä esimerkiksi osana tiedusteluprosessiaan isojen datamäärien louhintaan sekä analysointiin. Tekoäly mahdollistaa valtion tukemille ryhmille myös sen, että ne ovat kyenneet kehittämään sellaisia haittaohjelmia, joiden ei tarvitse käyttää komentopalvelimia niin aktiivisesti. Sen takia myös kiinnijääntі on epätodennäköisempää. Samaan aikaan tällaiset autonomiset haittaohjelmat osaavat piiloutua paremmin. Haastava tekijä valtiollisten uhkatoimijoiden tekoälypohjaisissa haittaohjelmissa on myös se, että niitä on vaikeampi kytkeä tiettyyn tekijään. Valtiollisten toimijoiden attribuointi on jo entisestään haasteellista ja tekoäly muuttaa asetelmaa entistä mutkikkaammaksi. Valtiollisten toimijoiden takana on lisäksi monetaarisia resursseja yleensä käytössä huomattavasti enemmän kuin rikollisryhmillä ja yksittäisillä tekijöillä. Sen ansiosta ne pystyvät rekrytoimaan tekoälyn parissa työskentele-

viä tutkijoita kehittämään tekoälypohjaisia kyberhyökkäyksiä. (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 20.)

Luvussa 4.2.4 mainittujen supertietokoneiden valjastaminen hyökkäyksellisiin toimiin kyberavaruudessa on merkittävä uhka, mutta Sharikovin (2018, s. 370) mukaan on epätodennäköistä, että tällaista tuomiopäivän laitetta tuskin on vielä millään valtiolla käytössä. On hyvä huomioida samalla, että edellä mainittu julkaisu on kuusi vuotta vanha. Tietotekniikan teknologisen kehityksen näkökulmasta se on pitkähkö aika. Nyt tilanne voi olla jo toinen.

4.2.7 Ennuste tulevaisuudesta

Tämän pääluvun lopuksi käsitellään Traficomin raportissa (2022, s. 22) esitettyä aikajanaa, joka kuvaa kehitystä lähitulevaisuudessa. Tekoäly voi tuoda paljon mahdollisuuksia kyberhyökkäysten kehittämiseen ja tehostamiseen, mutta keskeisenä ongelmana nähdään, ettei sitä ole kyetty vielä todistamaan. Kuviossa 8 on esitetty yksi ennuste tekoälypohjaisten hyökkäysten kehittymisestä. On hyvä huomioida, että kuvio on vuodelta 2022, joten siitä on tultu noin kaksi vuotta eteenpäin.

KUVIO 8 Kehitys lähitulevaisuudessa (*Tekoälyn mahdollistamat kyberhyökkäykset*, 2022, s. 22).

Tekoälyyn pohjautuvilla teknologioilla on Sharikovin (2018, s. 370) mukaan mahdollista korvata ihmishakkerit ohjelmistoilla, jotka kykenevät yhtäaikaisesti parantamaan omia vikojaan sekä haavoittuvuuksiaan. Samaan aikaan tällaisilla

ohjelmistoilla voi olla kyky edellä mainittujen kykyjen rinnalla etsiä uusia bugeja kohdejärjestelmästä ja hyödyntää niitä (Sharikov, 2018, s. 370).

Voidaan todeta, että kyberavaruudessa operoivien uhkatoimijoiden tekoälyn käytön aste omissa toimissaan on hyvin vaikea todistaa vedenpitävästi, ellei joissain tapauksissa mahdoton. Olisi kuitenkin välinpitämätöntä ja vastapuolen aliarvioimista ajatella, etteivät uhkatoimijat käyttäisi tekoälyä saavuttaakseen omia tavoitteitaan.

Tekoälyn käyttö aseena kyberavaruudessa on otettu kansainvälisessä valtioiden välisessä keskustelussa huomioon jo vuonna 2018. Tuolloin YK:n jäsenmaiden hallituksista koostuvan asiantuntijaryhmän kokouksessa käytiin keskustelua siitä, miten olemassa olevat kansainväliset lait soveltuvat tekoälyaseisiin. Autonomisista asejärjestelmistä, joilla on kyky tappaa, ei päästy yhteisymmärrykseen. Samaan aikaan kuitenkin ryhmän jäsenvaltioiden kesken vallitsi konsensus siitä, että taistelukentällä ihmisälykkyys on säilytettävä pääasiallisena päätöksentekijänä. Nähdään myös, että olisi tärkeää laatia uudet kansainväliset pelisäännöt, jotta vältettäisiin hallitsematon asevarustelukilpailu. (Sharikov, 2018, ss. 370–371.)

Se miten uhkatoimijat tulevat käyttämään tulevaisuudessa tekoälyä, on hyvin haastavaa ennustaa. Tässä alaluvussa esitettiin yksi skenaario, josta saa suuntaviivoja tähän kysymykseen. Yhdysvaltain asevoimien entinen kyberpäällikkö Paul Nakasone totesi vuonna 2018, että tekoäly tulee muuttamaan sekä puolustus- että hyökkäysoperaatioita viitaten vuonna 2016 järjestettyyn DARPA Grand Cyber Challenge -kilpailuun (*Trump's Pick for NSA/CyberCom Chief Wants to Enlist AI For Cyber Offense*, 2018). Kyseisen kommentin esittämisestä on nyt tultu noin kuusi vuotta eteenpäin ja voitaneen esittää varovainen arvio siitä, että näin todella on jo tapahtunut tai tapahtuu lähitulevaisuudessa.

Turtiaisen ym. (2023, s. 140) mukaan heidän artikkelinsa kirjoitushetkellä hyökkäykselliset tekoälyteknologiat ovat jo vahvasti esillä ja samanaikaisesti ne ovat jo kypsässä vaiheessa. Lockheed Martinin kybertappoketjua vasten tarkasteltuna esiin nousee selvimmin tiedustelu-, toimitus- ja asennusvaiheet, kun tarkastellaan tekoälyteknologioiden kypsyyttä, uusinta tekniikka edustavien töiden määrää sekä mahdollisten skenaarioiden kattavuutta. (Turtiainen ym., 2023, s. 140.)

Kybertappoketjun viitekehyksen näkökulmasta hyväksikäyttövaihe ja viimeinen eli varsinaisten toimien toteuttamisen vaihe ovat vielä alkutekijöissään ja Turtiainen ym. (2023, s. 140) toteavatkin, että nämä vaiheet edellyttävät huomattavasti enemmän tutkimusta ja yhtäältä panostuksia innovointiin. Nyt eikä tulevaisuudessa ei pidä myöskään unohtaa tai sivuuttaa, että hyökkäyksiä voidaan toteuttaa myös puolustavan osapuolen kyberturvallisuusjärjestelmissä käytettyjä tekoälyteknologioita vastaan. Tällaisissa järjestelmissä on havaittu heikkoja kohtia. Onkin hyvin todennäköistä, että uhkatoimijat tulevat hyödyntämään niitä. Tämä on tietenkin huolestuttava suuntaus. (Turtiainen ym., 2023, s. 140.)

Voidaan havaita, että tulevaisuuden haaste on myös, että tekoälyteknologioiden käyttöä on mahdoton rajoittaa vain lain oikealla puolella olevaan eettiseen käyttöön. Turtiainen ym. (2023, s. 140) esittävätkin tähän liittyen hyvin

kuvaavan analogian siitä, kuinka esimerkiksi keittiöveitsiä on helposti saatavilla erilaisista kaupoista.

He toteavatkin lopuksi, että tutkimusyhteisöllä ja kyberturvallisuusalan toimijoilla on kuitenkin mahdollisuuksia taistella uhkatoimijoiden tekoälyn väärinkäyttöä vastaan. Tähän yksi lääke on tekoälyteknologioiden kehitys, jonka lähestymistapana olisi suuntaus, jossa tekoälytekniikoita- ja malleja voitaisiin matalalla kynnyksellä tarkastaa, todentaa sekä selittää. Tähän yksi keino on selitettävä tekoäly. (Turtiainen ym., 2023, s. 140.)

4.3 Tekoäly kyberpuolustuksessa

Edellisessä alaluvussa tehtiin katsaus tekoälyn hyödyntämiseen uhkatoimijoiden ja hyökkäävän osapuolen näkökulmasta kyberturvallisuuden kentällä. Tässä alaluvussa on sen sijaan tarkoitus lähestyä tekoälyn hyödyntämistä kyberpuolustuksen näkökulmasta.

4.3.1 Nykytila

Kyberhyökkäykset ja kyberavaruudessa tapahtuva rikollisuus on muuttunut entistä monimutkaisemmaksi, mikä heijastelee suurena haasteena kyberpuolustukselle. Samoilla linjoilla ovat myös Wiafe ym. (2020, s. 598), joiden mukaan kyberpuolustuksessa käytettyjen menetelmien tulee olla entistä vankempia ja samaan aikaan älykkäämpiä. Kehittyneet kyberhyökkäykset vaativat puolustusteknologioilta sekä -mekanismeilta kykyä tehdä päätöksiä reaaliaikaisesti. Jotta tällaiseen tilanteeseen on mahdollista päästä ja saavuttaa nopea reagointikyky puolustavan osapuolen näkökulmasta, on tutkijoilla sekä kyberturvallisuusalan asiantuntijoilla iso rooli kehittämisen ja tutkimuksen osilta. Tähän lähestymistapaan nivoutuu vahvasti tekoäly ja sen käyttö kyberhyökkäysten torjunnassa. (Wiafe ym., 2020, s. 598.)

Kun tekoälyä lähdetään hyödyntämään puolustuksellisissa ratkaisuissa kyberturvallisuuden alalla, on hyvin tärkeää huomioida, että myös tekoälyjärjestelmät itsessään ovat alttiita hyökkäyksille. Ei siis ole turvallisuuden näkökulmasta lainkaan järkevää ajatella, että tekoäly olisi jonkinlainen viisasten kivi, jolla voidaan ratkaista ongelmat riskeittä. Onkin siis tärkeää tekoälymenetelmien implementoinnissa ottaa huomioon, että tekoälyn pitäisi olla samaan aikaan turvallista. Tekoälymallien ja etenkin selittävien tekoälymallien haavoittuvuuksia tullaan käsittelemään myöhemmässä alaluvussa. (Ali ym., 2023.)

Tänä päivänä organisaatioissa nojataan tietojärjestelmien osalta paljon pilvipalveluihin ja Wiafe ym. (2020, s. 598) nostavatkin esille konkreettisena esimerkkinä pilvi-infrastruktuurit ja niille ominaisen kolmikerrosarkkitehtuurin. Merkittävän haasteen tästä näkökulmasta muodostaa se tosiseikka, että jokaista kerrosta uhkaavat erityyppiset haavoittuvuudet. Nämä voivat juontaa juurensa ohjelmointivirheisiin tai ne voivat olla palveluntarjoajien aiheuttamia. Ongelmia kyberpuolustuksen näkökulmasta aiheuttaa myös datan epäyhtenäinen

käsittely. Se tekeekin kybervarkauksien- ja petosten tutkinnasta ja selvittämisestä entistä hankalampaa. Uhkatoimijoiden etu on samaan aikaan ympäri maailmaa hajautetut tietojärjestelmät, joiden ansiosta vihamieliset tahot voivat toimia mistä päin maailmaa tahansa. (Wiafe ym., 2020, s. 598)

Laskentatehojen kasvulla on ollut oma roolinsa siinä, kuinka tekoälyratkaisuista on kinostuttu kyberturvallisuuden alalla. Stanfordin yliopisto julkaisi vuonna 2019 AI Index -raportin, jonka mukaan suuren mittakaavan kuvien luokittelujärjestelmien kouluttamisessa pilvi-infrastruktuurin avulla tapahtuu merkittävää kehitystä. Raportin mukaan kouluttamiseen käytetty aika tulee pienenemään merkittävästi. Lokakuun 2017 ja heinäkuun 2019 välillä kolmesta tunnista 88 sekuntiin. Toinen merkittävä kehityskulku on tekoälyyn pohjautuvien lähestymistapojen laskentatehojen huomattava kasvu. Se tuplaantuu noin kolmen kuukauden välein, mikä tarkoittaa sitä, että vauhti on Mooren lakia kovempaa. (Zhang, Ning, ym., 2022, s. 1030.)

Gordon E. Moore ennusti 1965 julkaistussa Electronics -lehden numerossa julkaistussa ”Cramming more components onto integrated circuits” -artikkelissa, että vuoteen 1975 mennessä voisi olla mahdollista mahduttaa jopa 65000 komponenttia puolijohteisiin. Moore totesi, että komponenttien monimutkaisuus on tuplaantunut vuoden aikana ja lyhyellä aikavälillä vauhdin odotetaan kasvavan. Tähän hetkeen tultaessa vauhti on luonnollisesti hidastunut 1960-luvulla tehdystä ennusteesta. (Schaller, 1997, s. 53.)

Jos esimerkiksi viisitoista vuotta sitten organisaatiot saattoivat kohdata kyberhyökkäyksien skaalan, joka on laskettavissa kahden käden sormilla, niin tänä päivänä hyökkäykset ovat hyvin moninaisia ja kompleksisia. Kuten jo aiemmin mainittua, hyökkäykset ovat nykyään myös hyvin kehittyneitä, mikä osaltaan asettaa vaatimuksen tekoälyn ja älykkäiden agenttien käyttämiseksi, jotta näitä uhkia voidaan torjua. Nykytiedon valossa kyberturvallisuusratkaisujen pitääkin olla entistä älykkäämpiä, joustavampia sekä riittävän vahvoja, jotta ne kykenevät nykyuhkien havainnointiin ja samaan aikaan torjuntaan. (Wiafe ym., 2020, s. 598.)

Wiafe ym. (2020, s. 598) tuovat esiin samassa yhteydessä, että edellisen kappaleen kaltaiset vaatimukset edellyttävät puolustavalta osapuolelta tekoälytekniikoita kyberhyökkäysten riittävän tehokkaaseen valvomiseen ja torjuntaan. Yksi tärkeä edellytys tälle on, että tutkimuksen kentällä kuten myös alan asiantuntijoiden keskuudessa tunnetaan uusimmat tekoälymenetelmät kyberturvallisuuden saralla. (Wiafe ym., 2020, s. 598–599.)

Tieteen maailmassa informaatio- ja viestintäteknologioiden aloilla vallitsee konsensus ja yksimielisyys siitä, että tietoturva on ratkaisevan tärkeä asia, joka vaatii huomiota ja tutkimista entistä enemmän. Useat tutkimukset ovatkin keskittyneet esimerkiksi parannettujen teknologiaratkaisujen kuten haittaohjelmien ilmaisimien, tunkeutumisen havaitsemis- ja estämisjärjestelmien (engl. Intrusion Detection and Prevention System, IDPS), edistyksellisten palomuurijärjestelmien ja datan salausalgoritmien tutkimukseen. (Wiafe ym., 2020, s. 598–599.)

Wiafen ym. (2020, s. 599) mukaan on noussut esiin tutkimuksia ja niissä esitettyjä väitteitä, että tietoturvaan liittyvät kysymykset ovat hallittavissa tehokkaasti, jos ne keskittyvät ihmisen käyttäytymiseen. Samalla tutkimuskentällä on olemassa toinen laita, jonka mukaan ihmisen käyttäytymisen tutkiminen

ei riitä ratkaisemaan tietoturvaan liittyviä ongelmia. Konkreettisena esimerkkinä käytetään monissa organisaatioissa olevaa valtavaa datan määrää, joka vaatii automaation käyttöä merkittävissä määrin. Tästä juontuu turvallisuuteen liittyvien toimien osalta tarve yhdistää ihmiset, teknologia ja politiikkojen hallinta sekä samaan aikaan saavuttaa tasapaino näiden kaikkien kolmen välillä. (Wiafe ym., 2020, s. 599.)

Kun puhutaan perinteisitä kyberturvallisuuden teknisistä ratkaisuista, niin merkittäväksi ongelmaksi nousevat kiinteät ja muuttumattomat algoritmit ja fyysiset laitteet, kuten erilaiset sensorit ja ilmaisimet, joiden avulla voidaan havaita epätavallinen ja haitallinen toiminta tietojärjestelmässä. Tällaiset edellä kuvaillut konventionaaliset ratkaisut ovat tehottomia. Yhtenä reaalimaailman esimerkkinä voidaan käyttää ensimmäisen sukupolven viruksentorjuntamenetelmiä, jotka suunniteltiin tunnistamaan virukset siten, että niiden digitaaliset allekirjoitukset (engl. bit signature) skannataan. (Wiafe ym., 2020, s. 599.)

Lyhyesti kuvattuna edellä mainittu perinteinen tapa toimii siten, että lähdetään perusoletuksesta, jonka mukaan viruksella on kaikissa mahdollisissa tapauksissa sama rakenne sekä samoja bittikuvia eli bitit ovat samassa järjestyksessä. Tämän tyyppiset algoritmit ja allekirjoitukset ovat kiinteitä ja muuttumattomia. Tietoturvaohjelmistoja ja niiden kirjastoja sekä tietokantoja toki päivitetään allekirjoitusten osalta. Siitä huolimatta ei riitä, vaikka päivitykset tapahtuisivat tiheällä frekvenssillä, esimerkiksi päivittäin tai aina, kun laite on yhteydessä internetiin. Erityisesti laajimmat haittaohjelmat ovat niin kehittyneitä, että tällainen perinteinen ja vanhanaikainen lähestymistapa on yksinkertaisesti tehoton ja riittämätön. (Wiafe ym., 2020, s. 599.)

On esitetty lähestymistapoja, jotka ovat allekirjoituksettomia. Niillä on kyky havaita sekä lieventää haittaohjelmahyökkäyksien aiheuttamia tuhoja ja menetyksiä uudempien menetelmien avulla. Näihin lukeutuvat käyttäytymiseen liittyvät havainnot sekä erilaiset tekoälymenetelmät. Tällaiset lähestymistavat ovat väitetysti tehokkaampia kyberuhkia vastaan. (Wiafe ym., 2020.)

Wiafe ym. (2020, s. 599) esittävät ajatuksen siitä, että tämänkaltaisten modernimpien lähestymistapojen ja tekoälysovellusten kehittymisen myötä on tullut mahdolliseksi kehittää tehokkaita ja toimivia tietojärjestelmiä, joilla on kyky tunnistaa ja estää haitallista ja vihamielistä toimintaa kyberavaruudessa. Tällaisia menetelmiä on alettu hyödyntää perinteisempien menetelmien rinnalla, koska niiden avulla kyetään valvomaan ja ennaltaehkäisemään kyberavaruudessa tapahtuvia hyökkäyksellisiä toimia paremmin. (Wiafe ym., 2020, s. 599.)

On selvää, että tekoäly tarjoaa monenlaisia etuja myös kyberturvallisuuden kentällä. Oman haasteensa kuitenkin muodostaa se tosiseikka, että lähestymistapojen nopea kehitys ajaa tutkijat tilanteeseen, jossa voi olla hyvin vaikeaa tunnistaa, mikä tekniikka olisi tehokkain ja mitkä ovat vaikutukset kyberturvallisuuden näkökulmasta. (Wiafe ym., 2020, s. 599.)

Wiafen ym. (2020, s. 599) näkemys on hyvin kirkas, mitä tulee kyberturvallisuuskentän tutkijoihin sekä asiantuntijoihin siitä, että tekoäly on tuonut organisaatioiden tietoturvaan parannuksia. Samaan aikaan he toteavat kuitenkin, että tämänkaltaisilla väitteillä ei ole juurikaan vankkaa tieteellistä pohjaa, koska niitä ei ole empiirisesti perusteltu. Heidän mukaansa on havaittavissa, että useimmissa alan tutkimuksissa on nähtävissä kahdenlaisia tapauksia. En-

simmäisessä on osoitettu, että tutkimuksessa käytetty innovaatio on parempi kuin joukko olemassa olevia menetelmiä. Toisessa taas on tutkittu otos järjestelmiä ja sen jälkeen tehty arviointi suorituskyvystä, jossa on vastakkain tutkijoiden innovaatio ja otos olemassa olevista järjestelmistä. Huolestuttava haaste kaikissa tapauksissa on, että valikoitumisharhan ilmentyminen on korkeahko. (Wiafe ym., 2020, s. 599.)

Tämän takia tarve kootulle kirjallisuudelle on huomattava. Tarvitaan kirjallisuuden pohjalta yhteenvetoja, jotka koskevat alan ongelmia haasteita ja suuntaviivoja liittyen tulevaisuuden tutkimukseen. (Wiafe ym., 2020, s. 599.)

Wiafen ym. (2020, s. 605) tutkimuksen tulokset osoittavat, että SVM (engl. support vector machine) on algoritmina suosituin tunkeutumisen havaitsemis- ja torjuntajärjestelmissä. SVM-algoritmi on vakaa ja luotettava luokittelualgoritmi. Sen käyttö on kasvanut hyvin voimakkaasti vuosien 2013 ja 2018 välisenä aikana. SVM:n ehdottomana etuna on, että sen avulla voidaan ratkaista ongelmia, joissa näytteet ovat pieniä ja jotka ovat epälineaarisia. Näistä algoritmeista on muototutunut kyberturvallisuuden alalle standardinomaisia huippuluokan hyvin tunnettuja algoritmeja, mitä tulee tunkeutumisen estämis- ja havaitsemisjärjestelmiin (engl. Intrusion Detection and Prevention System, IDPS). (Wiafe ym., 2020, s. 605.)

Toinen suosittu tekniikka IDPS-järjestelmissä sekä haittaohjelmien, virusten, ja tietojenkalastelumenetelmien käsittelyssä on ensemble-tekniikka. Näitä tekniikoita on käytetty useissa tutkimuksissa, mutta kuitenkin vähemmän salauksen, palvelunestohyökkäysten, kuvantamisen ja captchan puolella. (Wiafe ym., 2020, s. 605.)

Ensemble-tekniikoilla ja -menetelmillä tarkoitetaan sellaista kokonaisluokittelijaa, joka yhdistää joukon muita luokittelijoita. Tavoitteena on siis saada aikaiseksi parempi luokittelija tästä yhdistelmästä. Ensemble-menetelmän avulla on tarkoitus imitoida sen kaltaista ihmisluontoa, joka pyrkii ennen päätöksentekoa tekemään harkintaa useampien mielipiteiden pohjalta. Ensemble-algoritmin on havaittu parantavan suorituskykyä esimerkiksi tarkkuuden ja väärien positiivisten (engl. false positive) osalta. (Menahem, 2009, s. 1483.)

Wiafe ym. (2020, s. 605) esittävät, että olisi tärkeää tutkia miksi ensemble-menetelmiä ei juurikaan käytetä salauksen, sertifioinnin, palvelunestohyökkäysten, kuvantamisen ja captchan osilta. Samaan aikaan he korostavat sen menestyksen muilla kyberturvallisuuden alueilla (Wiafe ym., 2020, s. 605).

Wiafe ym. (2020, s. 605) nostavat esiin kiinnostavan havainnon liittyen algoritmeihin, jotka ovat olleet aiemmin suositumpia ja parempana pidettyjä. Tällaisiin algoritmeihin lukeutuu esimerkiksi AdaBoost, jota on pidetty aiemmin luotettavana algoritmina (Wiafe ym., 2020, s. 605).

AdaBoost (engl. adaptive boosting) on tunnetuin buustausalgoritmi, jonka ideana on käyttää samojen harjoitusnäytteiden painotettuja versioita, kun toisena vaihtoehtona on valita harjoitusnäytteiden osajoukko satunnaisesti. AdaBoostin etu on, että se ei tarvitse isoa datamäärää harjoitteluun, koska sen algoritmi käyttää toistuvasti eri painotettuja versioita samoista harjoitusnäytteistä. (Shahraki ym., 2020, s. 5.)

Viimeistä edellisessä kappaleessa mainitun havainnon mukaan AdaBoost on eduistaan huolimatta menettämässä suosiotaan. Vaikka se onkin vankka ja

luotettava algoritmi, niin sen saama huomio näyttää jäävän pieneksi kyberturvallisuuden haasteita ratkaistaessa. (Wiafe ym., 2020, s. 605.)

Iso kuva tekoälyn suhteen kyberturvallisuuden alalla on kuitenkin sellainen, että tekoälysovellusten tuomat hyödyt ovat olleet merkittäviä ja ne ovat olleet menestyksekkäitä. Tutkimuksissa on havaittu, että uudemmat järjestelmät ovat olleet monella osa-alueella parempia verrattuna vanhoihin. Näihin osa-alueisiin listautuu esimerkiksi tarkkuus, energiatehokkuus, ennustetarkkuus, laskennallisen monimutkaisuuden väheneminen ja tekoälymallien koulutusaikojen lyheneminen. (Wiafe ym., 2020, s. 605.)

4.3.2 Kuinka hyötyä tekoälystä kyberturvallisuusratkaisuissa?

Luvussa 4.1. käsiteltiin tekoälyä yleisellä ja korkeammalla tasolla. Tässä on tarkoitus tehdä lyhyt katsaus siihen, miten tekoälystä voitaisiin hyötyä kyberturvallisuuden alalla ja hyödyntääkin parasta aikaa. Kuten aiemmin käsiteltyä, tekoälyllä tarkoitetaan tietojenkäsittelytieteiden osa-aluetta, joka yleisimmin painottaa älykkäiden koneiden luomista. Niillä on kyky ajatella ja toimia ihmisenkaltaisesti (Sarker ym., 2021, s. 2).

Tänä päivänä tekoälyä hyödynnetään kyberturvallisuuteen liittyvien haasteiden ratkaisemisessa jo hyvinkin moninaisesti ja laaja-alaisesti. Konkreettisena esimerkkinä tunkeutumisen havaitsemis- ja estämisjärjestelmien osalta älykkääseen ongelmaratkaisuun voidaan hyödyntää suosittuja tekoälytekniikoita, kuten koneoppimista (engl. machine learning, ML), syväoppimista (engl. deep learning, DL), luonnollisen kielen käsittelyä (engl. natural language processing, NLP), tiedon esittämistä ja päättelyä (engl. knowledge represnting and reasoning, KRR) ja tietoon tai sääntöihin perustuvien asiantuntijajärjestelmien (engl. knowledge or rule-based expert systems, ES) mallintamista. (Sarker ym., 2021, s. 2).

Edellä mainittujen tekniikoiden hyödyntämisessä on vain luovuus rajana. Yleisimmin niitä on hyödynnetty kyberturvallisuuden ratkaisuissa esimerkiksi haitallisen toiminnan tunnistamiseen, petosten havaitsemiseen, verkkohyökkäysten ennustamiseen, pääsynvalvonnan hallintaan ja anomalioiden havaitsemiseen verkkoliikenteessä. (Sarker ym., 2021, s. 2.)

Tänä päivänä tekoälyjärjestelmät on koulutettu siten, että ne kykenevät havaitsemaan organisaatiota vastaan kohdennettuja mahdollisia kyberuhkia, tunnistamaan uusia hyökkäysvektoreita, sekä suojaamaan organisaation tietoja. On listattu kolme tärkeää hyötyä, joita tekoälypohjaiset työkalut kyberturvallisuuden alalle tuovat: (*What Is AI in Cybersecurity?*, ei pvm.)

- Suurten tietomäärien nopea analysointi,
- Kyky havaita haavoittuvuudet ja poikkeamat,
- Toistuvien prosessien automatisointi.

Tekoälyn hyödyntämiselle kyberturvallisuuden kentällä ei ole olemassa rajoja – tai ainakin mahdollisuudet ovat lähes tulkoot rajattomat. Muutamana konkreettisena esimerkkinä voidaan avata tekoälyn tuomia hyötyjä.

Ensimmäisenä uhkien havainnointi sekä reagointinopeus uhkia kohtaan yhdistettynä korkeaan tarkkuuteen on hyvin lähellä reaaliaikaisuutta. Luvussa 4.2. käsiteltiin tekoälyn tuomia hyötyjä hyökkääjän näkökulmasta, missä todettiin, että hyökkääjät voivat säästää aikaa ja resursseja automatisoimalla tehtävät, joihin on tarvinnut aiemmin käyttää ihmisresursseja ja manuaalista työtä. Samankaltaisia etuja on nähtävissä myös toisella puolella. Yksinkertaistettuna puolustavan osapuolen osalta tekoälyn avulla voidaan tehostaa kyberturvallisuustomintoja. Täten tietoturvan parissa työskenteleville asiantuntijoille vapautuu arvokasta aikaa ja resursseja tärkeämpiin tehtäviin. (*What Is AI in Cybersecurity?*, ei pvm.)

Kuten jo tässäkin tutkielmassa aiemmin mainittua, yksi kyberturvallisuutteen liittyvistä päätehtävistä ja -tavoitteista on pitää huoli siitä, että yritysten ja yksittäisten ihmisten yksityisyys sekä tiedot pysyvät turvattuina. Tämä on yksinkertainen päämäärä, mutta sitäkin monimutkaisempi toteuttaa. Tämän alle nivoutuu esimerkiksi sovellus-, verkko-, tieto- sekä operatiivinen tietoturva. Kyberturvallisuuden katsotaan täyttyvän, kun nämä kaikki edellä mainitut toteutuvat. Tämä on edellytys sille, että yritysten kasvulla on paremmat lähtökohdat. Samaan aikaan myös liiketoiminnan jatkuvuus on vankemmalla pohjalla. Jatkuvuuden kannalta tekoälyllä on myös merkitystä, ja se voi olla auttava tekijä sen edistämisen näkökulmasta. (Ali ym., 2023.)

Ali ym. (2023) ovat samoilla linjoilla kuin Wiafe ym. (2020) tekoälyn tuomista hyödyistä kyberturvallisuuden alalle. He nostavat esiin erityisesti koneoppimisen ja sen tuomat edut etenkin olemassa olevien uhkien ennaltaehkäisyyn, jonka johdosta voidaan päästä parempaan tilanteeseen organisaatioiden tietojen ja yksityisyyden suojaamisen näkökulmasta (Ali ym., 2023).

On hyvä huomioida, että tekoäly tuskin tulee kuitenkaan koskaan korvaamaan tietoturvan ammattilaisia, eli ihmisiä, kyberturvallisuuden saralla. Tekoäly ei ainakaan toistaiseksi kykene ihmisten tekemään luovaan ongelmanratkaisuun tai monimutkaisempien haasteiden ratkaisuun. Kuten aiemmin mainittua, se missä tekoälystä saadaan suuria hyötyjä, on valtavien ja kasvavien datamäärien analysoinnissa, toistuvien kaavojen tunnistamisessa sekä päätelmien luomisessa suurten tietoturvadataan perustuvien datamassojen avulla. (*What Is AI in Cybersecurity?*, ei pvm.)

Guembe ym. (2022) ja Wiafe ym. (2020) puhuivat perinteisistä ratkaisuista tunnistaa uhkia ja esimerkiksi viruksien tunnistamisesta sääntöpohjaisten kyberturvallisuusratkaisuiden avulla. Tällaisiin vanhanaikaisiin tietoturvaratkaisuihin peilautuvat myös esimerkiksi työkalut, joiden avulla saapuvaa verkkoliikennettä verrataan tietokantaan, jossa on tunnettuja uhkia tai haitalliseen koodiin liittyviä allekirjoituksia. Kun työkalu tunnistaa uhkan, se tekee sitä käyttävälle asiantuntijalle hälytyksen ja ehdottaa toimenpiteitä uhkan estämiseksi. Nämä allekirjoituksiin (engl. signature-based) perustuvat mallit ovat olleet kohtuullisen tehokkaita tunnettuja uhkia vastaan. On havaittu, että tällaiset menetelmät ovat riittämättömiä tuntemattomien uhkien ja esimerkiksi nollapäivähyökkäysten edessä. (*What Is AI in Cybersecurity?*, ei pvm.)s

Nollapäivähaavoittuvuuksilla tarkoitetaan sellaisia aukkoja ohjelmiin liittyvässä tietoturvassa, joita ei ole kyetty vielä havaitsemaan ohjelmien kehittäjien toimesta. Niitä ei täten ole myöskään pystytty vielä paikkaamaan tai kor-

jaamaan. Kun tällaisia haavoittuvuuksia havaitaan verkkorikollisten ja uhkatoimijoiden toimesta ja niitä hyväksikäyttäen toteutetaan kyberhyökkäyksissä, niin niistä käytetään englannin kielen termiä zero-day attack eli nollapäivähyökkäys. (*Mikä on kyberhyökkäys?*, ei pvm.)

Väärät positiiviset hälytykset (engl. false positive) ovat kyberturvallisuuden alalla datamäärien kasvaessa merkittävä riesa. Väärillä positiivisilla tarkoitetaan hälytyksiä, jotka ilmoittavat virheellisesti, että haavoittuvuus olisi olemassa. (”NIST SP 800-115”, 2020.)

Allekirjoituksiin perustuvat havaitsemismenetelmät ja tämän kategorian alle nivoutuvat kyberpuolustustyökalut omaavat lisäksi varjopuolen liittyen vääriin positiivisiin havaintoihin. Tällä on sellainen seuraus, että kyberturvallisuusalan ammatti laiset, jotka käyttävät näitä työkaluja, joutuvat väärien positiivisten perässä jahtiin, joka voi johtaa täysin harhapoluille ja täten vie turhaan arvokasta aikaa muilta tehtäviltä. Tämän ongelman lyömiseksi tekoälyn avulla voitaisiin löytää ratkaisuja. (*What Is AI in Cybersecurity?*, ei pvm.)

4.3.3 Industry 4.0

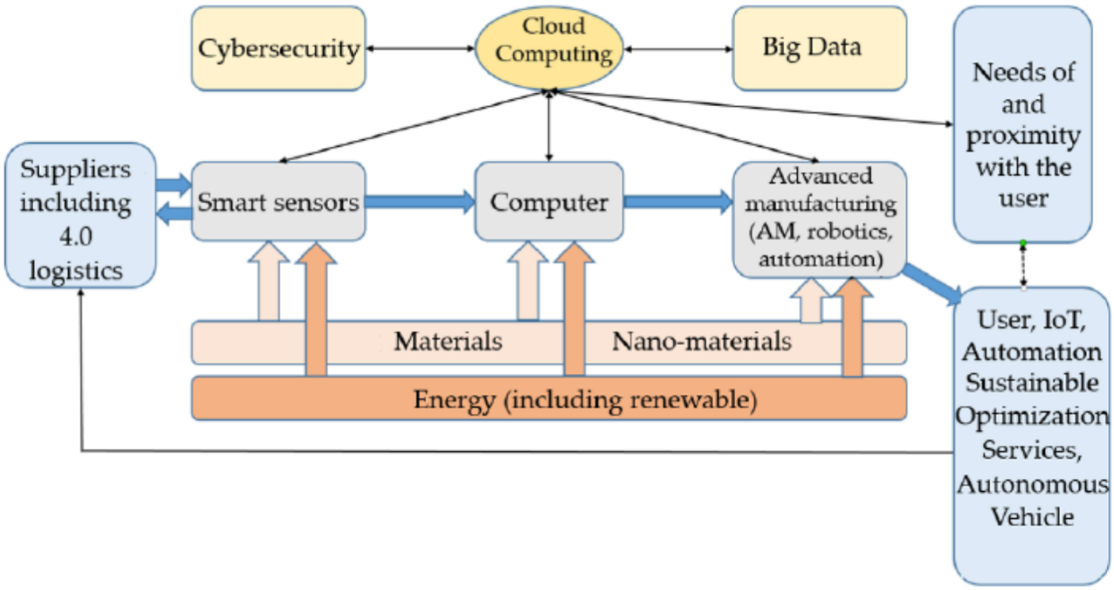
Kyberturvallisuuden, kuten monella muullakin alalla, kuuma puheenaihe ja lisää haasteita tuottava seikka on teollisuuden siirtymineen neljännen sukupolven aikakauteen. Neljännen sukupolven teollisuus (engl. industry 4.0) tuottaa tänä päivänä valtavia datamääriä, mikä johtaa entistä monimutkaisempiin tehtäviin kyberturvallisuuden ammattilaisten työpöydillä (Ali ym., 2023).

Neljännen sukupolven teollisuudella tarkoitetaan digitalisaation korkeaa tasoa, joka on nostanut tehokkuutta huomattavasti ja yhtäaikaisesti lisännyt joustavuutta eri teollisuuden aloilla. Näihin listautuvat esimerkiksi ympäristö-, kone-, energia- ja lääketieteelliset tekniikan alat monien muiden ohella. Industry 4.0:n merkittäviä osatekijöitä ovat esineiden internet (engl. Internet of Things, IoT), tekoäly, koneoppiminen, syväoppiminen, big data ja tiedonlouhintamenetelmät. Kaikki näihin edellä mainittuihin liittyvät sovellukset ovat osana lähes kaikkea automaatiota. (Gordan ym., 2023.)

Sarker ym. (2021, s. 2) käyttävät teollisuuden neljännestä sukupolvesta termiä neljäs teollinen vallankumous, joka on osuva kuvaus uudesta sukupolvesta monestakin lähestymiskulmasta tarkasteltuna. Heidän mukaansa tekoäly on neljännen sukupolven osalta avainteknologia. On mahdollista, että sillä on tärkeä rooli älykkäissä kyberturvallisuuden ratkaisuissa sekä hallinnoinnissa. (Sarker ym., 2021, s. 2.)

Sarker ym. (2021, s. 3) toteavat lisäksi, että kyberturvallisuus on avainasemassa teollisuuden neljännen sukupolven onnistumisen näkökulmasta. Tämä voidaan nähdä merkille pantavana lähestymistapana, koska neljännen sukupolven teollisuuden tietojärjestelmät ovat hyvin todennäköisesti vihamielisille toimijoille houkuttelevampia kohteita hyökätä ja taas esimerkiksi valtiollisia toimijoita kiinnostaa todennäköisesti entistä suuremmissa määrin vastustajavaltioiden teollisuusjärjestelmät kuten myös kriittisen infrastruktuurin kannalta merkittävät järjestelmät.

Kuviossa 9 (André, 2019, s. 3) on avattu neljännen sukupolven teollisuutta ja osia, joista se muodostuu. Wiafen ym. (2020) mainitsema pilviteknologioihin nojaaminen on läsnä myös tässä yhteydessä ja siihen liittyy läheisesti myös kyberturvallisuus, kuten kuviosta 9 voi havaita.



KUVIO 9 Industry 4.0 (André, 2019, s. 3).

Kyberturvallisuuden haasteita ratkoessa tekoälyteknologioilla, kuten koneoppimisella ja luonnollisen kielen käsittelyllä, on iso rooli datan keräämisessä ja sen analysoinnissa. Sen avulla kyetään yhdistämään ja rakentamaan kyberuhkatietoa, jonka avulla voidaan taistella tulevia kyberuhkia vastaan organisaatioissa. Tähän liittyy tiiviisti kyberuhkatiedustelu ja siihen liittyvät prosessit. (Ali ym., 2023.)

Euroopan unionin kyberturvallisuusvirasto (ENISA) onkin linjannut, että kyberuhkatiedustelun saralla tekoälyn tutkimusta on tarpeen jatkaa. Tutkimus edesauttaa ihmistyönä toteutettavien aikaa vievien manuaalisten työvaiheiden määrän vähentämistä kyberuhkatiedustelun analyyseissä. Tekoälyn tutkimisella voidaan vaikuttaa myös tietoturvariskien hallinta- ja lieventämissyklien analyysien validointiin koko syklin ajan. (Ali ym., 2023.)

Ali ym. (2023) toteavat, että tekoälyteknologioiden hyödyt kyberturvallisuuteen ylittävät sen aiheuttamat haitat. Tekoälyä myös tutkitaan samalla jatkuvasti enemmän sitä mukaa, kun kehitystä sen osalta tapahtuu. Tutkimuksen kasvu antaa suuntaviivoja merkittävälle teknologian kehitykselle kyberturvallisuuden ratkaisuissa. (Ali ym., 2023.)

4.3.4 Rajoitukset ja haasteet

Tekoälyn käyttöön kyberturvallisuuden ratkaisuissa liittyy kuitenkin useita haasteita ja ongelmia. Wiafe ym. (2020) huomauttavatkin, että vaikka nykyinen tutkimus on osoittanutkin lupaavia tuloksia, niin uhkia on samaan aikaan lu-

vassa ja niihin on reagoitava. Kyberturvallisuuteen liittyvät kysymykset ja ongelmat ovat skaalaltaan hyvin laajoja, ja se vaatiikin uusien tekoälyalgoritmien harkitsemista. Algoritmien avulla kyetään vastaamaan uusiin haasteisiin ja uhkiin. On ehdotettu, että olisi tarpeellista kasvattaa kiinnostusta ja huomiota liittyen hybridi- ja ensemble-luokittelijoiden käyttöön, jotta nykyisiä teknologioita saataisiin vietyä eteenpäin kyberturvallisuuden kentällä. (Wiafe ym., 2020.)

Vaikka tekoälyn hyödyntäminen kyberturvallisuuden ratkaisuissa on mullistanut alaa ja tulee jatkossakin tarjoamaan massiivisia hyötyjä, niin Sarker ym. (2021, s. 13) huomauttavat samalla, että tekoälypohjaiseen kyberturvallisuuteen liittyy useita tutkimusongelmia. Yksi merkittävimmistä haasteista on reaalimaailman tutkimuskysymysten ymmärtäminen sekä relevantin kyberturvallisuuteen liittyvän datan tutkiminen. Näistä edellä mainituista voidaan muodostaa laajempaa ymmärrystä ja rakentaa tietämystä toimia varten, jotka tapahtuvat tulevaisuudessa. (Sarker ym., 2021, s. 13.)

Datan keräämiseen liittyy kuitenkin itsessään myös ratkaistavia haasteita johtuen tietolähteiden laajasta skaalasta sekä niiden dynaamisesta luonteesta. Todellisen maailman datan keräämisen haaste kytkeytyy kaikkiin; strukturoituun-, puolistrukturoituun-, strukturoimattomaan- sekä metadataan. Tekoälypohjaisen kyberturvallisuuden ratkaisujen yksi päähaasteista onkin kyberturvallisuuteen liittyvän datan ymmärtäminen, kuten myös kerätyn datan integrointi ja hallinta tehokasta data-analyysia varten. (Sarker ym., 2021, s. 13.)

Toisena haasteena Sarker ym. (2021, s. 13) nostavat esille tehokkaan ja samalla älykkään ratkaisun löytämisen, jolla vastata kyberturvallisuuteen liittyviin ongelmiin. Kone- ja syväoppimismalleja on jo käytetty paljonkin tietoturvaan liittyvien ongelmien selättämiseksi, mutta parannettavaa löytyy vielä paljon. Tällaisiin malleihin lukeutuvat klusterointi, sääntöpohjaiset lähestymistavat, luokittelu, neuroverkot ja niin edelleen. Yhtenä apuvälineenä näiden kehittämiselle ehdotetaan kehittyneitä analytiikkakeinoja. Konkreettisena esimerkkinä voisi toimia hyökkäysmallien havainnointi aikasarjoissa, käyttäytymisanalyysi, tietoturvaan liittyvien piirteiden vaikutus mallintamisessa, tietoturvamallien yksinkertaistaminen ja optimointi ja niin edelleen. Nähdäänkin, että kehittyneillä analyysitekniikoilla, parannetuilla kone- ja syväoppimistekniikoilla sekä uusilla dataan perustuvilla algoritmeilla voisi olla vahva rooli. Huomioon otettavaa kuitenkin on, että tietoturvaongelmien luonteella on tähän vaikutus, kun tarkastellaan tulevia tutkimussuuntia. Lopulta tärkeimpänä nähdään kuitenkin se, että älykkään kyberturvallisuusjärjestelmän tärkeimpänä tehtävänä on muodostaa tehokas kyberturvan viitekehys. Sen tulee samalla tukea erilaisia tekoälytekniikoita. (Sarker ym., 2021, s. 13.)

Sarkerin ym. (2021) näkemyksiin peilattuna Ozkan-Okay ym. (2024) ovat huomattavasti optimistisemmilla linjoilla tekoälyn käyttämisen suhteen kyberturvallisuuden ratkaisuissa. He näkevät, että vaikka koneoppimismalleihin liittyy haasteita, niin mahdolliset hyödyt ovat kuitenkin merkittäviä. Tulevaisuuden näkymä on heidän mukaansa toiveikas sen suhteen, että teknologioiden kehittyessä tullaan todennäköisesti näkemään vieläkin tehokkaampia puolustuksellisia mekanismeja, mitä nyt on. Huomattava merkitys on eri instanssien ja organisaatioiden investointi tutkimukseen ja kehitykseen sekä samaan aikaan löytää ratkaisuja ongelmiin, jotka liittyvät mallien toteutukseen. Investointien

merkitys korostuu nimenomaan siinä, että koneoppimismallien täysi potentiaali saataisiin hyödynnettyä. (Ozkan-Okay ym., 2024, s. 251)

Myös Ozkan-Okay ym. (2024) ovat samoilla linjoilla tekoälyyn liittyvien haasteiden osalta datan näkökulmasta tarkasteltuna kuin Sarker ym. (2021). Jotta syväoppimismalleja saataisiin implementoitua tehokkaasti kyberturvallisuuden alalla, niin huomioonotettava ja tärkeä kulma on, että käytettävissä oleva data on korkealaatuista. Myös tekoälymallien tuottamien virheellisten tulosten mahdollisuus pitäisi minimoida ja tähän Ozkan-Okay ym. (2024, s. 251) nostavat yhdeksi konkreettiseksi esimerkiksi väärien positiivisten tulosten minimoimisen. Organisaatioilla on myös tärkeä tehtävä tekoälymallien implementointia harkitessa siitä, että pystyttäisiin varmistamaan, että käyttöönotettavat syväoppimismallit olisivat läpinäkyviä ja samaan aikaan selitettävissä. Molemmat edellä mainitut ominaisuudet ovat avaintekijöitä luottamuksen rakentumisen näkökulmasta sekä sen kannalta, että käytettyjen mallien tuottamat tulokset voidaan varmistaa. (Ozkan-Okay ym., 2024, s. 251)

Voidaan nähdä, että Zhangin, Ningin ym. (2022) näkemys on löyhästi Sarkerin ym. (2021) ja Ozkan-Okayn ym. (2024) välimaastosta, kun laitetaan vaakakuppeihin tekoälyn mullistava vaikutus kyberturvallisuuden alalle mahdollisuuksien ja optimistisen ajattelun sekä uhkakuvien ja haasteiden näkökulmasta. Zhang, Ning ym. (2022, s. 1039) toteavat hieman varovasti, että tekoälyllä voi olla tärkeä rooli kyberturvallisuuden alalla. Samassa he korostavat, että tekoälyä täytyy mukauttaa ja räätälöidä, jotta se istuisi paremmin kyberturvallisuusalan asettamiin vaatimuksiin. Alan nykyinen tutkimus on painottunut enemmän siihen, kuinka voidaan saavuttaa mahdollisimman nopea uhkien havaitseminen sekä miten parannettaisiin tekoälymallien havaintotarkkuutta ja datan louhintaominaisuuksia. (Zhang, Ning, ym., 2022, s. 1039.)

Tekoälymalleissa käytetyn datan korkeat vaatimukset nousevat esiin myös Zhangin, Ningin ym. (2022, s. 1042) toimesta. Jotta tekoälymallien harjoittelu voidaan saattaa valmiiksi, dataa tarvitaan lähtökohtaisesti massiivia määriä. Ennen kuin suurta datamassaa on käytössä, tekoälymalleja voidaan hyödyntää esimerkiksi sellaisiin operaatioihin, joissa datan kohinaa yritetään vähentää ja puuttuvia arvoja pyritään täyttämään. Valvottujen (engl. supervised) mallien käytössä ollessa, data on merkittävä manuaalisesti, mikä aiheuttaa luonnollisesti lisää ihmistyötä. Kyberturvallisuuden ratkaisuissa käytettyjen tekoälymallien suuren datan tarpeen takia haasteena on myös se, että mallien oikea-aikaisten päätösten tekemisen kyky ei ole kovin korkea. (Zhang, Ning, ym., 2022, s. 1042.)

4.3.5 Tekoälymallien läpinäkyvyyden puute

Tämän tutkielman kannalta tärkein huomio on Zhangin ym. (2022) esiin tuoma haaste, joka liittyy kyberturvallisuuden alalla käytettyjen tekoälymallien päätöksentekoprosessien läpinäkyvyyden puutteeseen. Merkittävä ongelma monestakin eri näkökulmasta on, että tekoälyn päätöksentekoprosessissa sen kehittämiseen osallistuvat ihmiset, mallien ohjelmoijat mukaan lukien, ovat tietämättömiä siitä, miksi tekoälymalli päätyy siihen tulokseen, mihin se päätyy. Näin ollen tekoälyn päätöksentekoprosessit kärsivät vakavasta ongelmasta eli pro-

sessin avoimuuden puuttumisesta. Täten tekoälymalleja kutsutaankin mustiksi laatikoiksi (engl. black-box). Tekoälymallit kykenevät luomis- ja itsensä kehittämisprosessien aikana toteuttamaan parametrien automaattista konfigurointia ja optimointia ilman, että ihminen puuttuu siihen. Organisaatioiden henkilöstöresurssien tehokkaan käytön näkökulmasta tämä on edullinen asia, mutta varjopuolena on, että tekoälymallien päätöksentekoprosesseja on vaikea selittää selkeästi. (Zhang, Ning, ym., 2022, s. 1042.)

Tekoälymalleissa huomioonotettavaa on, että vaikka ne voivat päästä korkeisiin tuloksiin tarkkuuden osalta testidataa käytettäessä, niin se ei kuitenkaan takaa sitä, että yhtä korkeisiin tarkkuusprosentteihin päästäisiin tilanteissa, joissa malli kohtaa tuntemattomia tapahtumia. Tilanteissa, joissa tekoälymallin tuotosta hyödyntävä ihminen ei ole samaa mieltä mallin antaman tuloksen kanssa, tullaan usein hankalaan tilanteeseen. Kun mallin päätöksentekoprosessia on vaikeaa tai jopa mahdoton selittää, niin ihmisten epäilys ja epävarmuus lisääntyy päätöksentekotuloksia kohtaan. (Zhang, Ning, ym., 2022, s. 1042)

Kaur ym. (2023, s. 24) puhuvat myös sen puolesta, että tekoälytekniikat ja -mallit vaativat kehittämistä kyberturvallisuuden alalla. Tarve kehittyneemmille tekniikoille on olemassa, jotta koko potentiaali saadaan hyödynnettyä, mitä tekoälyyn tulee. He listaavat yhdeksi ratkaisevaksi tekniikaksi selitettävän tekoälyn (XAI) todeten samalla, että sillä voi olla suuri vaikutus käytännönläheisten ja käyttökelpoisen tekoälyn kehittämiseen kyberturvallisuuden ratkaisuissa ja alalla. (Kaur ym., 2023, s. 24.)

Kaur ym. (2023, s. 24) esittävät väitteen, jonka mukaan hyvin ratkaisevaa kyberturvallisuuden alalla on, että siellä käytettyjen tekoälymallien algoritmien pitää olla avoimia. Eli tieto ja selitys siitä, miten ne päätyivät annettuun lopputulokseen, täytyy olla saatavilla. Tällä hetkellä merkittävä haaste onkin, että nykypäivän tekoälymallit eivät tällaista avoimuutta tarjoa, eivätkä selitä toimintaansa. Samaan aikaan ne tarjoavat suorituskykyä, jollaista ei olla ennen kyberturvallisuuden alalla nähty. Suorituskykyisyys nousee esiin esimerkiksi pimeän verkon tutkimuksissa sekä haavoittuvuuksien arvioinneissa. Toisessa vaakakupissa painaa hyvin voimakkaasti se tosiasia, että lähes kaikki edellä mainitun kaltaiset tekoälymallit ovat mustan laatikon malleja eli ne eivät ole tulkittavissa tai selitettävissä. (Kaur ym., 2023, s. 24.)

Kaur ym. (2023, s. 24) linjaavat, että tulevaisuuden tieteen kentällä pitäisikin vastata tekoälyn varjopuoliin ja sen vajavaisuuteen kyberturvallisuuden alalla kohdistamalla tutkimusta selitettäviin tekoälymalleihin. Tutkimuksen kannalta tärkeitä alueita ovat juuri tulkittavat ja selitettävät mallit ja miten ne voisivat parantaa käytettävien algoritmien suorituskykyä avaten samalla mustia laatikoita. Näin ollen päästäisiin tilanteeseen, jossa algoritmit olisivat hyväksyttävämpiä ja ennen kaikkea luotettavampia kyberturvallisuuden kentällä. (Kaur ym., 2023, s. 24.)

Seuraavassa luvussa siirrytään käsittelemään selitettävää tekoälyä, jossa näihin edellä mainittuihin haasteisiin pureudutaan hieman tarkemmin.

4.4 Selitettävä tekoäly (XAI)

Kuten tässä tutkielmassa on tuotu esille aiemmin, tekoäly ei suinkaan ole tuore keksintö. Tutkimusalueena se on alkanut kehittyä jo 1950-luvulla. Tieteen sekä käytännön maailmassa tekoäly on saavuttanut paljon huomiota ja on nähtävissä, että tämä trendi tulee jatkumaan kasvavissa määrin. (Meske ym., 2022, s. 53.)

Meske ym. (2022, s. 53) esittävät aiheellisen huomautuksen siitä, kuinka tämän päivän tekoälyssä käytetyt algoritmit ovat saavuttaneet pisteen, jossa ne ovat ylittäneet ihmisen tehtäväsuorituskyvyn monilla aloilla. Tekoälysovellukset ovat jo kyenneet päihittämään ammattipelaajat esimerkiksi pokerissa. Myös lääketieteen alalla tekoälyn avulla on onnistuttu saavuttamaan isoja askelia esimerkiksi rintasyövän tunnistamisessa. Tekoälypohjaiset sosiotekniset järjestelmät ovat siis ottaneet mittavia askeleita kehityksen osalta. (Meske ym., 2022, s. 53.)

Monella toimialalla on saavutettu piste, jossa tekoälypohjaisia ratkaisuja on implementoitu käyttötapauksiin, joilla voi pahimmissa skenaarioissa olla vakavia seurauksia ihmisten henkeen tai terveyteen. Konkreettisia esimerkkejä ovat lääketieteellinen diagnostiikka, ilmailussa käytetyt autopilotit, rikoksien uusimisen ennustaminen tuomioistuimissa, autonomisesti ajavat autot ja niin edelleen. Tavallisille ihmisille tekoäly näyttäytyy jokapäiväisessä elämässä esimerkiksi Google Homen tai Applen Sirin muodossa. (Meske ym., 2022, s. 53.)

Lienee kiistatonta, että tekoäly on tullut yhteiskuntiimme jäädäkseen, ja siitä on paljon hyötyä erilaisissa sovelluksissa. Samaan aikaan tekoälyn käytöllä ja sen tekemiin päätöksiin nojaamisella on kuitenkin varjopuolensa ja pahimmillaan suuretkin riskit. Tekoälymallien taustalla olevat algoritmit muuttuvat jatkuvasti monimutkaisimmiksi. Useimmissa tapauksissa tekoälymallit ovat niin kutsuttuja mustia laatikoita, kuten tässä tutkielmassa on tuotu aiemminkin esille. Se tarkoittaa, että niiden sisäiset oppimisprosessit ja niiden tuottamat tulokset eivät ole täysin tai lainkaan ymmärrettävissä tai tulkittavissa. Tekoälyn kehittämisessä niistä vastuussa olevat yhtiöt joutuvatkin jatkuvasti tasapainoilemaan tekoälymallien suorituskyvyn ja selitettävyyden välillä. Voi olla, että kompromissien hakemisella näiden kahden välillä saattaa olla massiiviset vaikutukset yksilöihin, yrityksiin ja kokonaisiin yhteiskuntiin. (Meske ym., 2022, s. 53.)

On selvää, että tekoälyyn nojaavien järjestelmien käyttöön liittyy moninaisia riskejä ja yhtenä merkittävimpänä ongelmana nähdään tekoälyyn liittyvä puolueellisuus ja vinoumat. Puhutaan myös automaatiovinoumasta, jolla tarkoitetaan ihmisen taipumusta luottaa liiaksi päätöksentekoon, joka on automatisoitu. Tämä voi taas johtaa siihen, ettei mustan laatikon tekoälymallien tapauksissa havaita niiden tekemiä virheitä. On myös huomionarvoista, että jos ihmisillä on ennakkoluuloja, niin yhtä lailla on myös tekoälymalleilla. Tällaiset ennakkoluuloiset järjestelmät voivat tahallisesti tai tahattomasti tuottaa puolueellisia päätöksiä. (Meske ym., 2022, s. 54.)

Koneoppimismalleille syötetty harjoitusdatana käytetyt tekstiaineistot sekä verkosta kerätyt kieliaineistot voivat sisältää inhimillisiä ennakkoluuloja. Se voi johtaa esimerkiksi sellaisen koneoppimismallin kehitykseen, jonka tekemät

päätökset ovat rasistisia tai sukupuoleen ennakkoluuloisesti suhtautuvia. Todellisessa maailmassa, jossa elämme, ja yhteiskunnat ovat eläneet koko ihmiskunnan historian ajan, on paljon ennakkoluuloja ihmisten ja yhteiskuntien välillä. Tästä juontuu se tosiasia, että harjoitusdata, jota koneoppimismalleissa käytetään, on aina jollain tavalla ennakkoluuloja sekä vinoumia sisältävää, mikä johtaa tekoälymallien vinoumiin. Reaalimaailman arkisia esimerkkejä on lukemattomia, joista yhtenä esimerkkinä voitaisiin mainita Applen kasvojentunnistusjärjestelmä, joka ei erottanut aasialaisia käyttäjiä tai Googlen tunneanalysaattori, joka kehittyi homofobiseksi antisemitistiseksi valvontajärjestelmäksi, joka kohdistui suhteettomasti vähemmistöihin kuuluviin asuinalueisiin. Kyseinen järjestelmä oli suunniteltu botiksi, joka keskustelee käyttäjien kanssa Twitterissä, mutta siitä kehittyi lopulta verbaalisesti loukkaava tekoälymalli. (Meske ym., 2022, s. 54.)

Tieteen kentällä ja tietojärjestelmiä tutkivien tutkijoiden yhteisössä selitettävyyden tutkiminen ei ole ennennäkemätön suuntaus. Jo 1980- ja 1990-luvuilla on tutkittu selitettävyyden tarpeellisuutta. Tuolloin tiedeyhteisössä käytiin keskustelua siitä, miten selitykset vaikuttaisivat käyttäjien parempaan ymmärrykseen tietojärjestelmistä. Ymmärryksen vaikutusta tutkittiin päätöksenteon tehokkuuteen ja tuloksellisuuteen sekä siihen, kuinka tietojärjestelmät koetaan hyödyllisyyden, helppokäyttöisyyden, tyytyväisyyden sekä luottamuksen näkökulmasta. Aiemmissa tutkimuksissa on todettu, että kokemattomammilla asiantuntijoilla on suurempi ja erilainen tarve selityksille, kuin kokeneemmilla asiantuntijoilla. (Meske ym., 2022, s. 55.)

Charmet ym. (2022, s. 790) toteavat myös tasapainoilun haasteen tekoälymallien suorituskyvyn sekä selitettävyyden kehittämisessä. Tutkijoiden kesken on viritetty keskustelua siitä, kuinka tarpeellista tekoälymallien olisi olla selitettäviä eettisistä ja yhtä lailla toiminnallisista näkökohdista tarkasteltuna. DARPA:lla (Defense Advanced Research Projects Agency) on selitettävään tekoälyyn keskittyvä ohjelma, jossa on nostettu esiin, että koneoppimismallien selitettävyys on kääntäen verrannollinen sen ennustussuorituskykyyn ja -tarkkuuteen. Kuten aiemmin tässä tutkielmassa käsiteltyä, syväoppimismallit ovat itsessään jo hyvin monimutkaisia ja niiden algoritmit ovatkin kaikista monimutkaisimpia, mikä tarkoittaa, että ne ovat myös kaikista hankalimmin selitettävissä. (Charmet ym., 2022, s. 790.)

Myös Islam ym. (2022, s. 2) ottavat kantaa suorituskyvyn ja selitettävyyden väliseen kysymykseen ja samalla he toteavat, että näiden kahden välisellä kompromissilla voi olla vaikutuksia yksilöihin, yrityksiin sekä yhteiskuntiin. He esittävät myös huolenaiheen liittyen neuroverkkojen monimutkaistumiseen. Samalla, kun malleista tehdään monimutkaisempia, niiden selitettävyys ja tulkittavuus heikkenee entisestään tai muuttuu jopa mahdottomaksi. Samaan aikaan niillä on käsittämättömiä kykyjä korkean ennustetarkkuuden suhteen. Neuroverkkojen kyvykkyys on johtanut siihen, että niitä käytetään mustista laatikoista huolimatta ennustusten tekemiseen, jotka voivat olla ratkaisevia ja niillä voi samaan aikaan olla suuria vaikutuksia. Yhtäältä mustan laatikon malleja käytetään kriittisessä päätöksenteossa ja yhteyksissä. Mustan laatikon tekoälymallien tekemät päätökset aiheuttavat ongelmien kärjistymistä päätöksien yhteydessä, joita ei voida perustella. Ne eivät siis ole rationaalisia tai niitä ei

voida selittää. Monilla aloilla on tänä päivänä ollut tehtävissä havaintoja, että tekoälyyn nojaavien järjestelmien kehittyminen ja korkea kyvykkyys on tehnyt ihmisen toiminnan tarpeettomaksi. Mitä enemmän tekoälyn tekemiin päätöksiin ja tuotoksiin nojataan, sitä enemmän on myös tarvetta ymmärtää sekä tulkita tekoälypohjaisten järjestelmien tekemiä päätöksiä ja ennusteita. Useissa sovelluskohteissa ja monilla toimialoilla selitykset tai tulkinnat ovat välttämättömiä. Tällaisia esimerkkejä ovat täsmälääketiede, autonomisten ajoneuvojen reaaliaikaiset päätökset tai laillinen tietoliikenne osana verkkoturvallisuutta. Tällaisissa kohteissa käytetty tekoäly ja sen tekemät päätökset voivat olla hyvinkin kriittisessä roolissa. Niillä voi olla valtava vaikutus jopa ihmishenkiin. (Islam ym., 2022, s. 2–3.)

4.4.1 Mitä on selittävä tekoäly?

Russel ym. (2022, s. 728) kuvaavat koneoppimismenetelmän prosessin, jonka ensimmäisessä vaiheessa malli kehitetään harjoitusdatan avulla. Toisessa vaiheessa valitaan hyperparametrit validointidatan avustamana ja lopullinen metriikka saadaan testidatan avulla. Tässä prosessissa onnistuminen ei Russelin ym. (2022, s. 728) mukaan kuitenkaan ole riittävä edellytys sille, että koneoppimismalliin voisi luottaa. Jos mallin kehittäjä ei voi luottaa siihen, eivät voi myöskään muut sidosryhmät, kuten sääntelyviranomaiset, lainsäätäjät tai käyttäjät (Russell ym., 2022, s. 728).

Koneoppimismallit saattavat olla jopa sovelluskehittäjille hankalasti lähestyttäviä ja jonkinlaisen sumuverhon peitossa. Sitä tekoäly ja koneoppiminen sen osa-alueena ovat varmasti myös monelle tavalliselle käyttäjälle.

Koneoppimisjärjestelmät ja -mallit ovat kuitenkin vain ohjelmistoja, joiden luottamusta voidaan rakentaa perinteisillä sovelluskehityksessä käytetyillä työkaluilla, joilla voidaan todentaa ja validoida mitä tahansa ohjelmistoa. Näihin työkaluihin listautuvat esimerkiksi lähdekoodin hallinta, joka pitää sisällään versiohallinnan, rakentamisen ja bugien seurannan. Toisena työkaluesimerkkinä voidaan käyttää testausta eli monenlaiset eri testit, joita ohjelmoinnissa ja sovelluskehityksessä käytetään mukaan lukien kuormitustestit, regressiotestit ja niin edelleen. (Russell ym., 2022, s. 729.)

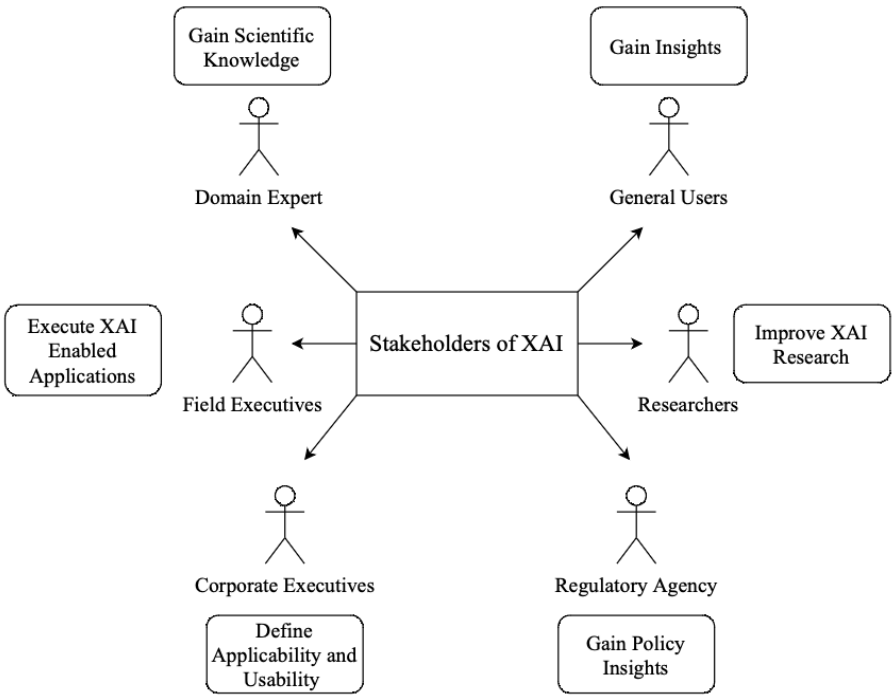
Russel ym. (2022, s. 729) painottavat näiden edellä mainittujen lisäksi selitettävyyden ja tulkittavuuden tärkeyttä koneoppimismalleille. Selitettävissä oleva tekoälymalli tiivistettynä on malli, joka auttaa lisäämään ymmärrystä, miksi tietty mallin tuottama tuotos saatiin aikaan jollain tietyllä syötteellä (Russell ym., 2022, s. 729).

Russel ym. (2022, s. 729) erottelevat selitettävyyden ja tulkittavuuden heidän terminologiassaan. Tulkittavuuden nähdään juontuvan varsinaisen mallin tutkiskelusta ja selitettävyys tuotetaan erillisen prosessin avulla. Eli tekoälymalli johon tutkiminen kohdistuu, voi olla hankalasti ymmärrettävissä oleva mustan laatikon malli, mutta selitysmoduuli voi tehdä tiivistyksen siitä, miten malli toimii (Russell ym., 2022, s. 729).

Selitettävyys onkin siis yksi ja varsin hyvä tapa rakentaa luottamusta tekoälymalleja kohtaan. Tärkeää on myös huomioida, että EU:n tietosuoja-asetus

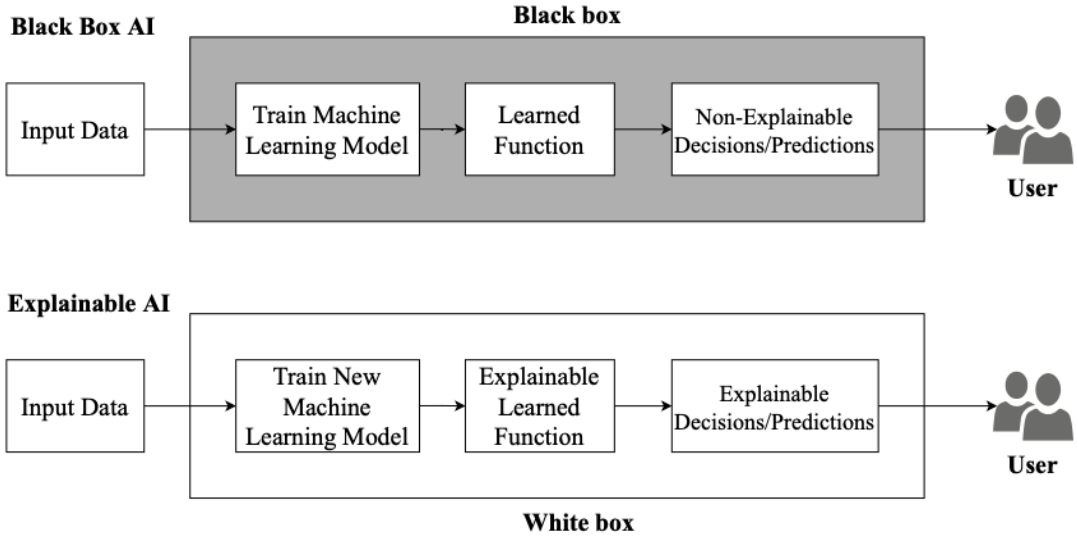
(General Data Protection Regulation, GDPR) edellyttää, että järjestelmät antavat selityksiä tekemilleen päätöksille ja toiminnalleen (Russell ym., 2022, s. 730).

Selitettävästä tekoälystä on muotoutunut oma tutkimusalueensa 2000-luvulla. Tähän yksi vaikuttava tekijä on ollut tekoälyyn liittyvien sidosryhmien jatkuvasti suuremmaksi kasvanut huoli tekoälymalleja kohtaan. Sidosryhmiin, jotka kytkeytyvät selitettävään tekoälyyn ovat kyseisen toimialan asiantuntijat, tutkijat, yritysten päättäjät, alan johtajat, sääntelyvirnaomaiset sekä käyttäjät.



KUVIO 10 Selitettävän tekoälyn sidosryhmät (Islam ym., 2022, s. 5).

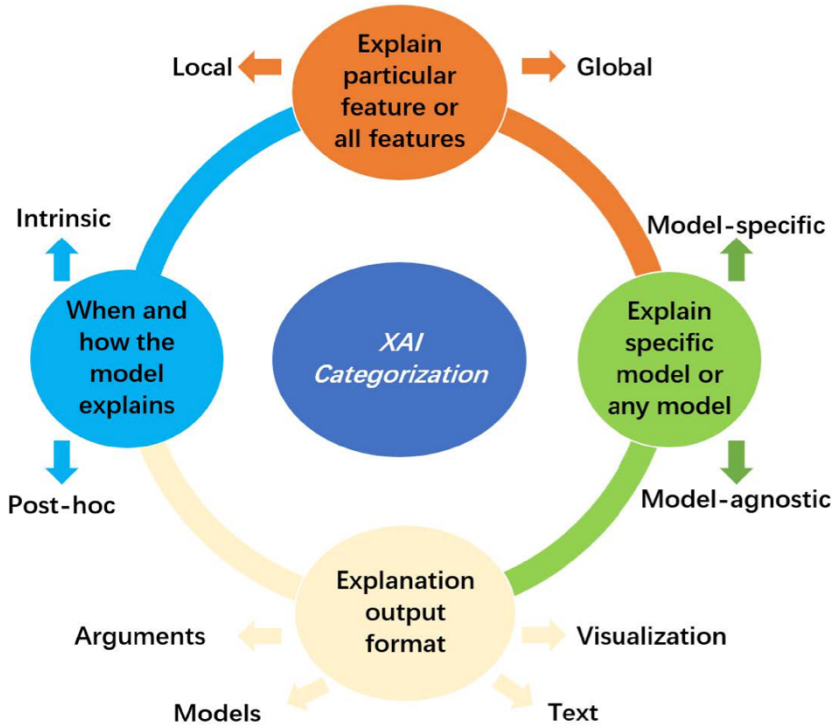
Selitettävän tekoälyn implementointiin erilaisissa tietojärjestelmissä liittyy varmasti organisaatiokohtaisia tavoitteita, mutta keskeisin tavoite on luoda tekoälymallin ennalta määriteltyjä päätöksiä avaavia selityksiä. Selitysten pitää olla sellaisia, että niitä pystyy ymmärtämään, vaikkei tekoälymallista olisi teknistä ymmärrystä. Yhtäältä selitysten täytyy olla luotettavia. Kuviossa 11 on kuvattu tekoälyn ja selitettävän tekoälyn keskeisimmät erot. Islam ym. (2022, s. 4) kiteyttävät selitettävän tekoälyn koostumuksen selitysprosessista sekä ihmisen ymmärryksestä sitä kohtaan.



KUVIO 11 Perinteisen mustan laatikon ja selitettävän tekoälymallin eroavaisuudet (Islam ym., 2022, s. 5).

4.4.2 Selitysmallit

Selitettävän tekoälyn selitysmallit voidaan kategorisoida useasta eri näkökulmasta. Kategorisointi on kuitenkin haastavaa johtuen siitä, että selitysmallien välisiä päällekkäisyyksiä voi esiintyä. Samaan aikaan jokin tietty selitystekniikka voidaan liittää yhteen tai vaihtoehtoisesti useampaan luokkaan. Kuviossa 12 on havainnollistettu eri selitettävän tekoälyn kategoriat.



KUVIO 12 Selitettävän tekoälyn kategorisointi (Zhang, Hamadi, ym., 2022, s. 111).

Mallikohtaiset (engl. model-specific) ja malliagnostiset (engl. model-agnostic) mallit tarkoittavat selitysmallien jaottelua sen mukaan, onko kyseinen malli sovellettavissa vain jollekin tietylle mallille. Mallikohtaisesta selitystyökalusta konkreettinen esimerkki on Graph Neural Network Explainer (GNN), joka on menetelmä, joka soveltuu GNN-pohjaisten mallien ennustusten tekemiseen. Malliagnostisia selitysmenetelmiä voidaan taas soveltaa lähes mihin tahansa koneoppimismallin selittämiseen. Ne toimivat käytännössä siten, että analysoidaan ominaisuuksien (engl. feature) syötteitä (engl. input) ja tuloksia (engl. output). Tällaisilla menetelmillä ei ole pääsyä mallien sisäiseen informaatioon, kuten painotuksiin tai rakennetietoihin. Esimerkkejä malliagnostista menetelmistä ovat SHAP-työkalut, Saliency Map sekä Gradient-weigheted Class Activation Mapping. (Zhang, Hamadi, ym., 2022, s. 111)

Kuviossa 12 esiintyvässä mallissa vasemmalla laidalla ovat paikalliset (engl. local) ja globaalit (engl. global) selitysmenetelmät. Tässä jaottelussa paikallisella selitettävyydellä tarkoitetaan sitä, että tähän kategoriaan menevät selitysmallit kuvaavat järjestelmän kykyä osoittaa käyttäjälle, miksi jokin tietty päätös on tehty. LIME ja SHAP voidaan konkreettisina esimerkkeinä niputtaa tämän alle. Globaaleilla selitysmenetelmillä tarkoitetaan sellaisia työkaluja, joissa selitettävyys viittaa oppimisalgoritmin selittämiseen kokonaisuutena. Yksi globaali menetelmä on Global Attribution Mapping (GAM). Sen avulla muodostetaan selityksiä neuroverkkojen ennusteiden toiminnasta. (Zhang, Hamadi, ym., 2022, s. 111)

Russel ym. (2022, s. 737) käyttävät selitysmallien jaottelussa jakoa kahteen eri luokkaan. Ensimmäinen on tekoälyjärjestelmien suunnittelijoita varten ja toinen on suunnattu tekoälyn käyttäjille. Samtani ym. (2022) tekevät myös selitysmallien jaon kahteen:

- Sisäiset (engl. intrinsic),
- Jälkikäteiset (engl. post-hoc) selitysmallit.

Myös kuvioon 12 sisältyvät sisäiset lähestymistavat käyttävät esimerkiksi päätöksentekosääntöjä, huomiomekanismeja ja päättelypolkuja. Tämän tyyppiset mallit toimivat tekoälymallin suorituksen aikana. Toiseen luokkaan asettuvat jälkikäteiset selitysmallit taas tarjoavat nimensä mukaisesti selityksiä sen jälkeen, kun kohteena oleva tekoälymalli on tehnyt suorituksensa ja antanut tuloksensa tai päätöksensä. Post-hoc -tekniikoita voidaan hyödyntää lähes mihin tahansa tekoälymalliin. (Samtani ym., 2022.)

Paljon käytettyjä selitystekniikoita ovat LIME-järjestelmät (engl. Local Interpretable Model-gnostic Explanations), jotka ovat menetelmiä, jotka rakentavat tulkittavia lineaarisia malleja. Sen sijaan aiemminkin mainittu toinen paljon käytetty selitysmetodi, SHAP (engl. Shapley Additive exPlanations) nojaa Shapley-arvoihin kunkin ominaisuuden (engl. feature) kontribuution määrittämiseen. (Russell ym., 2022, s. 737.)

LIME:n toiminta pohjautuu riippumattomuuteen siitä, mitä malliluokkaa käytetään. Se rakentaa tulkittavissa olevan mallin, joka on usein päätöspuu tai lineaarinen malli. Malli, jonka LIME rakentaa on tulkittavan mallin approksimaatio, ja siten se tulkitsee lineaarista mallia luodakseen selityksiä. Näissä

selityksissä lopulta tuodaan esiin, kuinka tärkeä mikin ominaisuus on tekoälymallin tuottaman tuotoksen kannalta. LIME toimii siis siten, että se käsittelee kohteena olevaa koneoppimismallia mustana laatikkona. Se kokeilee mallia erilaisilla satunnaisilla syöttöarvoilla, jonka perusteella se rakentaa tietokokonaisuuden. Sen pohjalta tulkittava malli voidaan lopulta muodostaa. Tämän tyyppinen lähestymistapa on hyvin soveltuva strukturoituun dataan. (Russell ym., 2022, s. 730.)

SHAP-selitysmalli puolestaan juontaa juurensa peli teoriasta (engl. game theory) ja se pyrkii päättelemään kunkin ominaisuuden kontribuution kohteena olevan tekoälymallin antamaan päätökseen siten, että se kuvaa ominaisuudet pelaajina pelissä, jossa on luotu erilaisia yhteenliittymiä. SHAP-selitysmallien tavoitteena on siis selittää mallin ennustetta laskemalla jokaisen ominaisuuden osuus ennusteeseen. Menetelmässä lasketaan siis Shapley-arvot. Ominaisuusarvoja voitaisiin kuvailla pelaajiksi, ja Shapley-arvot antavat käyttäjälleen tuloksen sekä ennusteen ominaisuuksien kesken. SHAP voidaan jaotella KernelSHAP:iin ja Tree-SHAP:iin. KernelSHAP laskee tapauksen x osalta kunkin ominaisuuden panoksen ennusteeseen. Tree-SHAP taas on SHAP:n modifikaatio puupohjaisille koneoppimismalleille kuten päätöspuille (engl. decision tree) ja satunnaismetsille (engl. random forest). (Spyrou & Kappatos, 2023, s. 863.)

Russel ym. (2022, s. 730) huomauttavat, että selityksiin ei pidä kuitenkaan luottaa sokeasti. Yksinkertainen selitys voi aiheuttaa vääränlaista turvallisuuden tunnetta. Samalla he toteavat, että yleensä käytetyksi tavaksi ratkaista jokin ongelma, valikoituu koneoppimismalli perinteisen käsin koodatun ohjelman sijaan. Tämä johtuu siitä, että usein ongelmat, joita yritetään ratkaista ovat jo luonnostaan monimutkaisia eikä perinteisen ohjelman kirjoittaminen sen ratkaisemiseksi ole mahdollista. Täten pitää ottaa huomioon, ettei jokaiselle koneoppimismallin tekemälle ennusteelle löydy mitään yksinkertaista selitystä. Täydellistä mallia ei siis ole olemassa. (Russell ym., 2022, s. 730.)

Selitysmenetelmien tuotokset (engl. output) liittyvät läheisesti siihen, kuinka menetelmät kategorisoidaan. Selityksen muodolla on merkittävä vaikutus siitä näkökulmasta, kuka mallia hyödyntää ja käyttää. Tekstipohjaisia selitysmenetelmiä käytetään esimerkiksi luonnollisen kielen prosessointiin (engl. natural language processing, NLP). Visualisoidut selitysmenetelmät taas puolestaan ovat laajemmin käytettyjä. Niitä hyödynnetään esimerkiksi neuroverkkojen selittämisessä. Kuviossa 12 selitysten tulokset on havainnollistettu ympyrän alalaidassa keskellä. (Zhang, Hamadi, ym., 2022, ss. 111–112)

4.4.3 XAI kyberturvallisuuden kentällä

Kuten luvussa 4.2 havaittua, uhkatoimijat pyrkivät olemaan aina puolustavaa osapuolta edellä käytettyjen tekniikoiden ja työkalujen osilta. Näin ollen vihamieliset toimijat ovatkin implementoineet tekoälyä vahvasti omissa käyttötapauksissaan jo vuosia. Zhang, Hamadi ym. (2022, s. 105) toteavat saman, ja samalla he nostavat esille kyberturvallisuustutkijoiden kiinnostuksen tekoälyyn perustuvia lähestymistapoja kohtaan, joihin lukeutuvat erityisesti kone- ja syväoppiminen.

Zhang, Hamadi ym. (2022, s. 105) mainitsevat, että tekoälytekniikoilla ja nimenomaisesti kone- ja syväoppimisalgoritmien avulla voidaan saavuttaa merkittäviä tuloksia monissa kyberturvallisuuden sovelluksissa. Tähän joukkoon lukeutuu esimerkiksi tunkeutumisen havaitseminen (engl. Intrusion Detection System, IDS), petosten havaitseminen, roskapostin suodatus, bottiverkkojen havaitseminen sekä haittaohjelmien tunnistus. Kuten tässä tutkielmassa useassa yhteydessä on todettu tekoälyratkaisuja käsiteltäessä, suorituskyvyn ja selitettävyyden välillä joudutaan usein tasapainottelemaan. Samoilla linjoilla ovat myös Zhang, Hamadi ym. (2022, s. 105). Edellä mainitut kyberturvallisuuden sovellukset tuloksiensa vakuuttavuudesta huolimatta sisältävät usein virheitä, joiden seuraukset voivat olla vaikutukseltaan huomattavia. (Zhang, Hamadi, ym., 2022, s. 105)

Selitettävyyden ja tulkittavuuden kustannuksella on kyetty toteuttamaan hyvinkin tarkkoja tekoälymalleja kyberturvallisuuden ratkaisuissa, mutta se on johtanut käytettyjen mallien monimutkaisuuteen ja vaikeaan hahmotettavuuteen. Luvussa 4.4.1 mainittiin EU:n tietosuoja-asetukseen (GDPR) sisällytetty vaatimus järjestelmien selitettävyydestä. Tämän nostavat esiin myös Zhang, Hamadi ym. (2022, s. 105) todeten samalla, että kyberturvallisuusjärjestelmiin luottaminen ja uskominen edellyttää niissä käytettyjen tekoälymallien läpinäkyvyyttä ja tulkittavuutta. Tutkimuksen puute on kuitenkin selitettävien tekoälymallien käytöstä kyberturvallisuuden alalla ilmeinen, vaikka tutkimuksia on alkanut nousta esiin viime aikoina jonkin verran. (Zhang, Hamadi, ym., 2022, s. 105).

Aiemmin selitettävää tekoälyä ja kyberturvallisuutta on tutkittu kyllä erikseen, mutta ei juurikaan yhdessä. Edellä mainittujen alojen välillä on kuitenkin noussut esiin yhtymäkohtia. Olemassa olevissa tutkimuksissa keskitytäänkin lähinnä vain tekoälysovelluksien, kuten kone- ja syväoppimisen, analysointiin kyberturvallisuuden alalla. (Zhang, Hamadi, ym., 2022, s. 105)

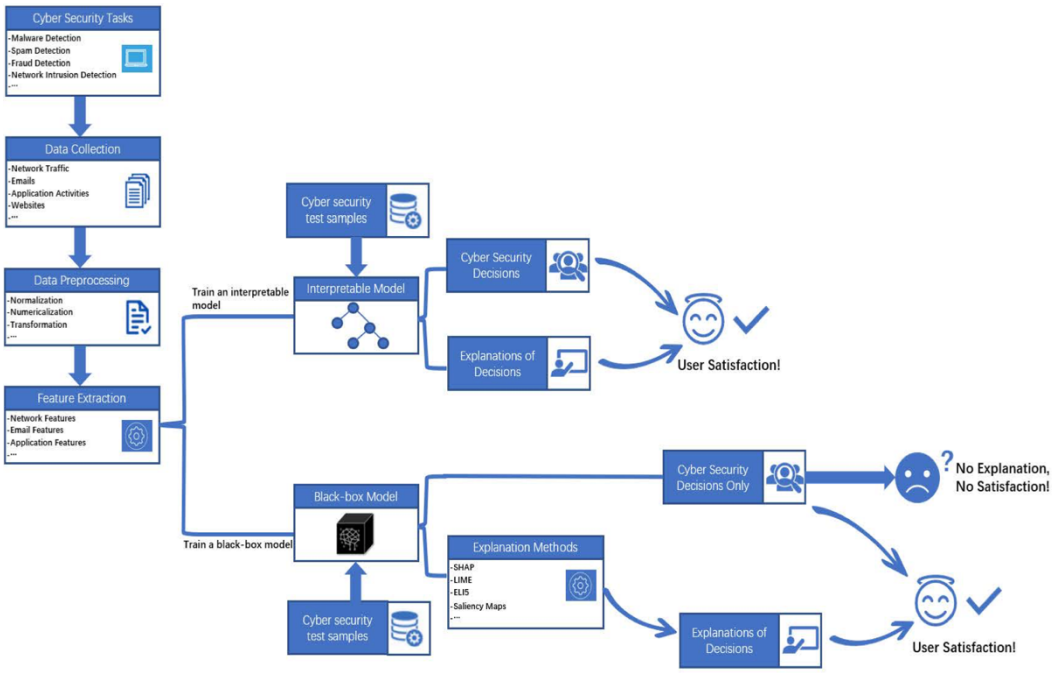
Myös Charmet ym. (2022, s. 791) toteavat tekoälyn olevan suuri osatekijä kyberturvallisuuden tutkimusympäristössä. He mainitsevat samalla, että on tärkeää kiinnittää huomiota siihen, kuinka tekoälyä käytetään asianmukaisesti kyberturvallisuuden kontekstissa. Tärkeys tästä näkökulmasta liittyy siihen, että oikeanlainen käyttö on haastava tehtävä. (Charmet ym., 2022, s. 791.)

Tässä pro gradu -tutkielmassa on käsitelty aiemmin muutamia aihealueita, joissa tekoälyä hyödynnetään kyberturvallisuuden ratkaisuissa ja sovelluksissa. Charmet ym. (2022, s. 791) nostavat esiin tekoälyn hyödyntämisalueiden esimerkkeinä verkkoturvallisuuden, tietokoneiden tietoturvan ja mobiiliympäristöt. Tekoälyn avoimuudella ja luottamuksella on tärkeä rooli kyberturvallisuuden kontekstissa Charmetin ym. (2022, s. 791) mukaan sen vuoksi, että selitettävyys vähentää pahantahtoisuutta, joka liittyy tekoälyjärjestelmiin.

4.4.4 XAI:n soveltaminen kyberturvallisuudessa

Tässä alaluvussa tehdään tiivis katsaus, miten selitettävää tekoälyä on tähän mennessä hyödynnetty kyberturvallisuuden kontekstissa ja kentällä. Kuviossa

13 on havainnollistettu selitettävän tekoälyn käyttöä kyberturvallisuuden tehtävissä verrattuna mustan laatikon tekoälymalleihin.



KUVIO 13 Selitettäviä tekoälyä hyödyntävien sovellusten kaaviokuva (Zhang, Hamadi, ym., 2022, s. 114).

Kuviossa 13 työnkulku etenee kiteytetysti alkaen kyberturvallisuuteen liittyvien tehtävätyyppien määrittelystä. Tyyppejä voivat olla esimerkiksi haittaohjelmien, verkkoon tunkeutumisen tai roskapostin havaitseminen. Seuraavat vaiheet sisältävät datan keräyksen, joka kohdistuu esimerkiksi verkkoliikenteeseen tai sähköposteihin. Tätä seuraa poiminta, joka kohdistuu merkittäviä ominaisuuksia edustaviin piirteisiin, jotka sittemmin syötetään koulutusdataksi tekoälymalleille. Testinäytteitä otetaan analysoitavaksi sen jälkeen, kun mallit on koulutettu. Näytteiden perusteella voidaan muodostaa päätöksiä. Kuvion 13 loppupäässä on kuvattu, kuinka käyttäjät voivat saada käytetystä tekoälymallista selityksiä. Kuten aiemminkin käsiteltyä, malli voi muodostaa selitykset suorituksensa aikana tai vaihtoehtoisesti voidaan käyttää esimerkiksi LIME- ja SHAP -tekniikoita jälkikäteen muodostettaviin selityksiin. Huomioitavaa on, että kuviossa 13 esitetty kaavio on kuitenkin vain yleisluontoinen selitettävien tekoälymallien käytön työnkulku kyberturvallisuuden alalla. Yksityiskohdissa voi olla täten varianssia eri tehtävissä. (Zhang, Hamadi, ym., 2022, s. 113)

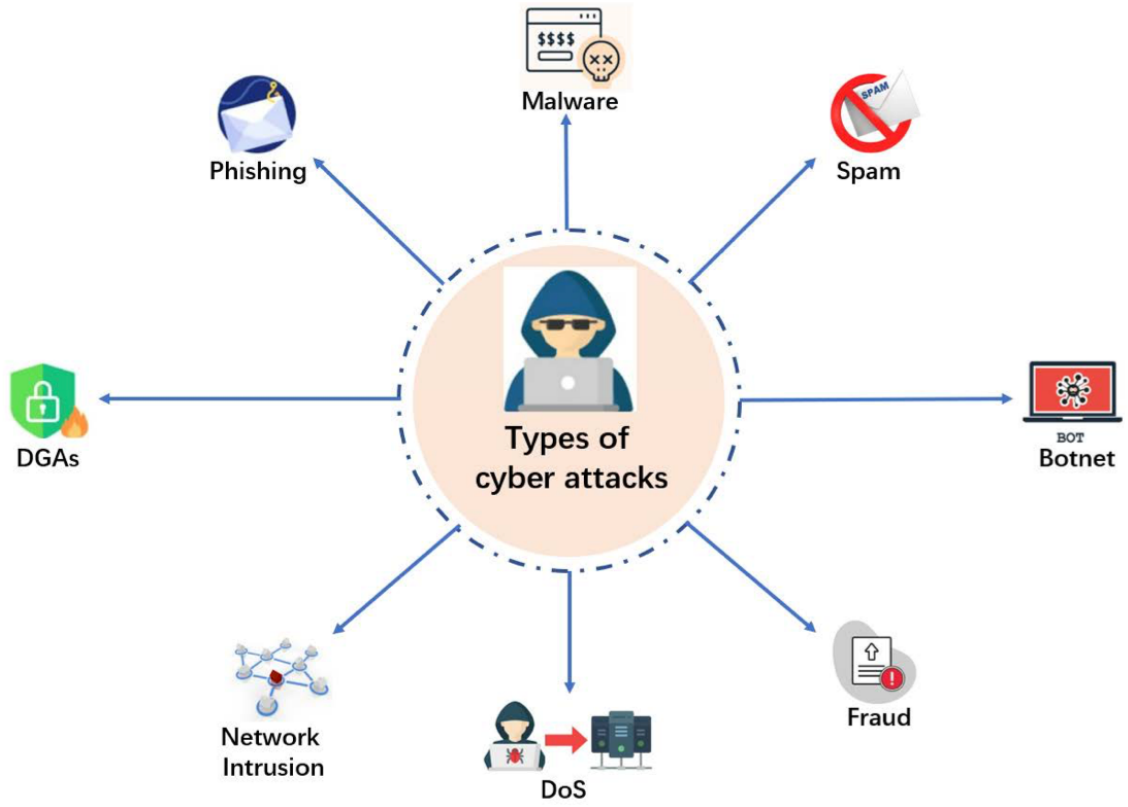
Zhang, Hamadi ym. (2022, s. 113) painottavat kyberturvallisuuden tutkimuksessa käytettävän datan laatua. Sillä on huomattava vaikutus selitettävien tekoälymallien tekemiin päätöksiin. Näihin malleihin lukeutuvat mukaan koneja syväoppimismallit. Kyberturvallisuusdataa voidaan kerätä hyvinkin yksinkertaisesti ja helposti esimerkiksi kaappaamalla verkkoliikennettä Wireshark -ohjelmistolla, jolla kaapataan verkkopaketteja. Tämän tyyppisiin menetelmiin liittyy kuitenkin haasteita, joista tärkeimpänä on vähäinen datan määrä ja sen kapeus. (Zhang, Hamadi, ym., 2022, s. 113)

Zhang, Hamadi ym. (2022, s. 113) ehdottavatkin, että on hyödyllisempää ja järkevämpää käyttää tekoälymallien koulutukseen kyberturvallisuusalalle kohdennettuja datakokonaisuuksia (engl. cyber security datasets). Tähän on monia syitä, joista yhtenä tärkeimpinä ajansäästö ja sen myötä vapautuvat resurssit tutkimiseen. (Zhang, Hamadi, ym., 2022, s. 113)

Kyberturvallisuuden alalla käytetyt XAI-sovellukset voidaan jaotella Zhangin, Hamadin ym. (2022, s. 114) mukaan kolmeen eri pääkategoriaan. Ensimmäinen ryhmä on puolustuksellisen selitettävien tekoälyjärjestelmien hyödyntäminen kyberhyökkäyksiä vastaan. Toinen ryhmä liittyy selitettävien tekoälymallien mahdollisuuksiin kyberturvallisuuden alalla ja kolmas kyberuhkiin ja puolustusmenetelmiin, jotka liittyvät selitettäviin tekoälymalleihin. (Zhang, Hamadi, ym., 2022, s. 114)

XAI-sovelluksia voidaan hyödyntää monia eri kyberhyökkäyksiä vastaan. Kuviossa 14 (Zhang, Hamadi, ym., 2022, s. 117) on kuvattu yleisimmät kyberhyökkäyksien tyypit. Käsitellään seuraavaksi kahden konkreettisen esimerkin avulla, miten XAI-sovelluksista voidaan saada apua taistelussa kyberuhkia vastaan.

Ensimmäisenä tarkastellaan selitettävien mallien roolia haittaohjelmien (engl. malware) havainnoinnissa. Perinteiset tekniikat, joita haittaohjelmien havainnoinnissa ja analysoinnissa käytetään, ovat todella työläitä ja vaativat runsaasti ajallisia resursseja. Esimerkiksi erilaisia syväoppimismalleja hyödynnetäänkin paljon haittaohjelmien havaitsemistyössä, koska niiden avulla saavutetaan parempi suorituskyky ja tarvitaan vähemmän resursseja. Käytetyt menetelmät nojautuvat täten neuroverkkojen käyttöön, jotka ovat mustia laatikoita eli eivät selitä toimintaansa. Siitä johtuen tutkimuskentällä tutkijat ovatkin ottaneet käyttöön vaihtelevasti erilaisia selitettävän tekoälyn lähestymistapoja. Niiden avulla tekoälyyn pohjautuvista haittaohjelmien havaitsemiseen käytetyistä järjestelmistä on kyetty saamaan selitettävämpiä ja läpinäkyvämpiä. Taustalla on tavoite, joka liittyy siihen, että luotettava haittaohjelmien havaitsemisjärjestelmä olisi edelleen yhtä luotettava, kun se implementoidaan uuteen ympäristöön käytettäväksi. Haittaohjelmien havaitsemisjärjestelmiä voidaan selittää useilla eri tavoilla. Yhtenä esimerkkinä toimii identifiointi, joka kohdistuu merkittävimpiin paikallisiin piirteisiin. Tunnistaminen voi tarjota selityksiä päätöksille, joita havaitsemisjärjestelmä tekee. Tällaisten päätösten nähdään olevan arvokkaita. (Zhang, Hamadi, ym., 2022, s. 117)



KUVIO 14 Yleisimmät kyberhyökkäyksien tyypit (Zhang, Hamadi, ym., 2022, s. 117).

Toisena esimerkkinä käytetään selitettävien tekoälymallien käyttöä verkkoon tunkeutumisen havaitsemisjärjestelmissä (engl. Network Intrusion Detection System, NIDS). Verkkoon tunkeutumisella tarkoitetaan luvatonta tunkeutumista esimerkiksi jonkin organisaation verkkoon. NIDS-järjestelmien tehtävänä on havaita epätavallinen tai vihamielinen käyttäytyminen verkossa. NIDS-järjestelmien rakentamiseen on haittaohjelmien havaitsemisjärjestelmien tavoin käytetty syvä- ja koneoppimismalleja, jotta järjestelmistä saataisiin tehokkaampia. Selitettävyyden implementoimista on harkittu tuotavan osaksi NIDS-järjestelmiä alan asiantuntijoiden toimesta. Tällä pyritään siihen, että mustan laatikon tekoälyjärjestelmiä saataisiin kehitettyä kestävämmiksi. XAI:n käyttöä tällä lähestymiskulmalla on jo kokeiltu. (Zhang, Hamadi, ym., 2022, s. 122)

Konkreettisena esimerkkinä on esitelty selitettävään tekoälyyn pohjautuva tunkeutumisen havaitsemisjärjestelmä (IDS), jonka tavoitteena on havaita haitallinen liikenne, joka tunkeutuu verkkoon. Tällaiseksi liikenteeksi voidaan katsoa esimerkiksi hajautetut palvelunestohyökkäykset (Distributed Denial of Service, DDoS). Kyseisessä järjestelmässä on hyödynnetty neuroverkkoja sekä puumalleja. Tästä järjestelmästä tehtiin havaintoja, että se vähensi konvoluutiokerrosten määrää neuraalisessa työssä. Tällä oli täten vaikutus siihen, että mallin selitettävyyttä saatiin parannettua, ja samaan aikaan tarkkuus ja suorituskyky eivät kärsineet. Tapauksessa käytettiin neuroverkon ennustustulosten käsittelemiseksi XGBoost -koneoppimiskirjastoa. Käsitellyt tulokset syötettiin sen jälkeen LIME- ja SHAP -järjestelmiin, joiden avulla voitiin muodostaa lisäselityksiä. (Zhang, Hamadi, ym., 2022, s. 122)

5 TULOKSET

Tässä luvussa avataan tutkimuksen tulokset. Kuten aiemmin mainittua, tutkimus toteutettiin kvalitatiivisena tutkimuksena. Tutkimusmenetelmäksi valikoitui laadullinen sisällönanalyysi. Seuraavaksi esitellään analyysin jokaisen vaiheen tulokset.

Aineistoa lähdettiin tiivistämään redusoimalla sitä. Alkuperäisilmauksista muodostettiin pelkistettyjä ilmaisuja kuten tämän tutkielman toisessa luvussa esiteltyä. Taulukossa 9 on vielä avattu kyseinen prosessi. Kyseisessä taulukossa on esitelty osa yhden analysoitavana olleen artikkelin redusoinnista. Taulukoissa 10–14 on redusoinnin tulokset eli kaikki pelkistetyt ilmaisut, joita analysoitavasta aineistosta muodostettiin.

Klusteroinnin, eli ryhmittelyn, tuloksena syntyneet alaluokat on esitetty taulukoissa 15–21. Alaluokkia muodostui yhteensä 22 kappaletta. Alkuperäisilmaisuja oli redusointivaiheessa yhteensä 203 kappaletta, joten aineistoa saatiin tiivistettyä hyvin.

Aineiston abstrahointi on kuvattu taulukoissa 22 ja 23. Alaluokista muodostettiin yhteensä kuusi yläluokkaa. Edelleen yläluokista muodostettiin lopulta kolme yhdistävää luokkaa. Yhdistävät luokat olivat:

- XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa,
- perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan,
- XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla.

TAULUKKO 9 Redusointiprosessi.

Tutkimuskysymys	
1. Miten selittäviä tekoälymalleja voitaisiin hyödyntää kyberuhkien havaitsemisessa?	
2. Voidaanko selitettävillä tekoälymalleilla saavuttaa korkeampaa luotettavuutta kyberturvallisuuden alalla?	
3. Kuinka selitettävät mallit soveltuisivat organisaatioille, jotka vaativat korkeaa laatua ja tarkkuutta käytettäviltä tietojärjestelmiltä?	
Aineisto/dokumentti	Tekijät
Phishing Email Detection Using Persuasion Cues Phishing-sähköpostin havaitseminen suostutteluvihjeiden avulla	Rohit Valecha , Pranali Mandaokar, and H. Raghav Rao
Tyyppi	Vuosi
Tieteellinen artikkeli	2022
Alkuperäisilmaukset	Pelkistetyt ilmaukset
Tutkimuskysymyksiin vastaamiseksi luomme kolme koneoppimismallia, joissa on relevantteja suostutteluvihjeitä (gain persuasion cues), tappiovihjeitä (loss persuasion cues) ja yhdistettyjä voitto- ja tappiovihjeitä (gain and loss persuasion cues), ja vertaamme estimaatteja perusmalliin, jossa ei oteta huomioon suostutteluvihjeitä. Tulokset osoittavat, että kolme phishing-havaintomallia, joissa on relevantteja suostutteluvihjeitä, ovat F-pistemäärältään noin 5-20 prosenttia paremmat kuin perusmalli, joten ne ovat luotettavia menetelmiä phishing-sähköpostin havaitsemiseen.	- Käytetyn havaitsemismallin luottamuksen kasvattaminen - Suostutteluvihjeiden käyttö phishing-sähköpostien havaitsemisessa
Tällainen tutkimus on hyödyllistä, koska suostutteluvihjeiden syvempi ymmärtäminen voi auttaa suunnittelemaan tehokkaita vastatoimia phishing-sähköpostien havaitsemiseksi ja estämiseksi.	- Suostutteluvihjeiden tehokkuus vastatoimien kehittämisessä
Ongelma kärjistyy entisestään, kun tietojenkalastelijat vaikuttavat uhrien reaktioihin ja pitävät samalla yllä aitoa ja laillista ulkoasua, jolloin suodattimien on vaikea luokitella tietojenkalastelusisältöä vilpilliseksi ja käyttäjät ovat edelleen alttiita tietojenkalasteluviesteille. Näin ollen sähköpostiviestien merkitseminen phishing-sisällöksi on yksi tehtävä, jota monet tietoturva-analyytikot tietoturvaoperaatiokeskuksissa joutuvat tekemään manuaalisesti. Siksi tarvitaan automaattisia, selitettäviä lähestymistapoja, jotka voivat tarjota arvokkaita tietoja phishingin havaitsemiseen sekä phishing-tietojen laadulliseen ja määrälliseen analysointiin.	- Manuaalisen ihmistyön kuormittavuus - Tarve selitettäville lähestymistavoille
Tätä tarkoitusta varten tässä asiakirjassa tarjotaan toimintakelpoista tiedustelutietoa kehittyvistä phishing-hyökkäyksistä (eli kyberuhkien tiedustelutietoa) käyttämällä selitettävyyttä, joka perustuu yhteiskuntatieteellisestä ja psykologisesta kirjallisuudesta peräisin oleviin teoreettisiin näkökulmiin, phishing-hyökkäysten tehokasta havaitsemista varten. Siinä tutkitaan erityisesti suostutteluvihjeiden tehokkuutta phishing-sähköpostin havaitsemisessa tai phishing-sähköpostisuodattimien suunnittelussa.	- Selitettävyyden hyödyt phishingiin liittyvässä kyberuhkatiedustelussa
Tutkijat ovat hiljattain tutkineet suostutteluvihjeiden roolia phishing-alttiudessa ja uhrien käyttäytymisessä. Laajennamme tätä tutkimusta mittaamalla suostutteluvihjeiden tehokkuutta phishing-sähköpostin havaitsemisessa.	- Suostutteluvihjeet selitettävyyttä tukemassa
Tällainen tutkimus on tärkeää, koska syvällisempi ymmärrys phishing-viestintekijöiden taktiikoista voi antaa tietoa sellaisten tehokkaiden vastatoimien suunnittelusta, joilla puututaan suoraan turvallisuusongelmiin ja jotka voivat auttaa phishing-viestien havaitsemisessa ja estämisessä. Tämä tutkimus voi auttaa ymmärtämään tietojenkalasteluviestien havaitsemisjärjestelmää laskennallisten ja inhimillisten tekijöiden avulla. Teoreettisen näkökulman hyödyntäminen voi auttaa tuottamaan selitettäviä malleja, jotka voivat tarjota tulkintoja mustan laatikon malleista (NISTIR 83123).	- Vastatoimien kehittäminen - Ymmärrys hyökkääjän taktiikoista - Teoreettisen näkökulman tuoma hyöty selitettävien mallien tuottamisessa - Black-box -mallien tulkinta
Viimeaikaisessa kirjallisuudessa on kuitenkin tuotu esiin NLP:n haittapuolia phishing-sähköpostin havaitsemisessa, sillä synonyymien käyttöä ja lauserakennetta on vaikea löytää NLP:n avulla. Siinä on myös todettu luokittelijaluokan ongelmia yleisesti, koska koneoppimismenetelmät perustuvat pääasiassa sähköposteissa edustettujen piirteiden tuottamiseen, mikä voi vaatia raskasta manuaalista työtä ja alaan liittyvää asiantuntemusta	- Perinteisten ML-mallien vaatima manuaalinen työ - Vanhat mallit raskaita
Vaikka monissa tutkimuksissa on analysoitu phishing-sähköpostiviesteissä esiintyviä suostutteluvihjeitä ja joissakin tutkimuksissa on osoitettu, että suostuttelu voidaan havaita automaattisesti, on vain rajoitetusti tutkittu suostutteluvihjeiden tehokkuutta phishingin havaitsemisessa. Tämän ongelman ratkaisemiseksi luomme integroidun kehyksen, jonka avulla voidaan kehittää tietokoneella mitattavia ominaisuuksia, jotka kuvaavat voittoa (vastavuoroisuuden, johdonmukaisuuden ja miellyttävyyden kautta) ja tappiota (tappion, vakavuuden ja välittömyyden kautta), jotta phishing-sähköpostiviestejä voidaan havaita.	- Käytetyn havaitsemismallin adaptoituminen ihmismäiseen käytökseen

TAULUKKO 10 Redusoinnin tulokset osa 1.

- Käytetyn havaitsemismallin luottamuksen kasvattaminen - Suostutteluvihjeiden käyttö phishing-sähköpostien havaitsemisessa	- DL mallien heikko tulkittavuus monimumtkaisuuden vuoksi
- Suostutteluvihjeiden tehokkuus vastatoimien kehittämisessä	- ML mallien käsitteleminen useimmiten selitettävinä malleina - Mallin tuottaman tuloksen selitettävyys
- Manuaalisen ihimistyön kuormittavuus - Tarve selitettäville lähestymistavoille	- Selitettävyyden vaikutus mallin luotettavuuteen
- Selitettävyyden hyödyt phishingiin liittyvässä kyberuhkatiedustelussa	- XAI:n harvinaisuus tietoturvasovelluksissa - Tutkimuksen puute selitettävyyden osalta
- Suostutteluvihjeet selitettävyyttä tukemassa	- XAI mallien tutkimuksen puute WF-hyökkäyksissä - XAI:n hyödyntäminen identiteetin ja yksityisyyden suojaamisessa
- Vastatoimien kehittäminen - Ymmärrys hyökkääjän taktiikoista - Teoreettisen näkökulman tuoma hyöty selitettävien mallien tuottamisessa - Black-box -mallien tulkinta	- XAI käsitteen historian tuoreus - XAI menetelmien jakautuminen kahteen kategoriaan
- Perinteisten ML-mallien vaatima manuaalinen työ - Vanhat mallit raskaita	- XAI sovellusten käyttö yleisesti - XAI kyberturvallisuuden alalla alustavasti
- Käytetyn havaitsemismallin adaptoituminen ihmismäiseen käytökseen	- Miten XAI:ta tutkittu kyberturvallisuuden alalla tähän mennessä
- Aiemmin käytettyjen DL mallien vaikea ymmärrettävyys kalasetun havaitsemisessa - DL mallit ovat black-box malleja - DL mallien selitettävyys edistynyt - Selitettävyys lisää ymmärrystä mallin tuottamista ratkaisuista	- XAI mallien tutkimuksen puute WF-hyökkäyksissä - XAI:n käyttö puolustusstrategioissa
- Selitettävän havainnointimallin tehokkuus	- XAI:n tarkoitus mallien käytöksen tutkimisessa.
- Selitettävän havainnointimallin tehokkuus - Selitettävän havainnointimallin luotettavuus	- Ehdotetut XAI-algoritmit - Algoritmien soveltuvuus erilaisiin malleihin - LIME ja Saliency Map -mallien soveltaminen WF-hyökkäyksein tutkimiseen
- Selitettävän havainnointimallin tehokkuus - Selitettävän havainnointimallin luotettavuus - Johdonmukainen suorituskyky luokittelussa - Yleistettävyys	- ML ja DL -mallien vaikea selitettävyys - Mallin tulkinnan helppous - Mallin käyttäytymisen selitettävyys
- Selitettävän havainnointimallin suorituskyky	- XAI analyysin haasteet
- Parempi havainnointikyky	- LIME:n käyttö selittämisessa
- Aikaisemman tutkimuksen puute	- XAI menetelmät piirteiden havaitsemisessa - XAI menetelmien arviointi
- Ristiriita, mitä selitettävyys on tieteellisen ymmärryksen näkökulmasta - Selittämisen ymmärtäminen - Poikkitieteellisen tutkimuksen tarve	- ROAR -metriikan käyttö ja ominaisuudet
- Selitettävän havainnointimallin suorituskyky	- ROAR selitettävien menetelmien arvioimisessa - ROAR:n XAI menetelmien tehokkuuden arvioinnissa
- Selitettävän mallin toiminta verrattuna perinteisiin luokittelumalleihin	- LIME:en pohjautuvan XAI-puolustustekniikan tehokkuus - RF-mallin väärinluokitteluaste
- Suositus organisaatioiden strategioiden kehittämiseksi selitettävän AI mallin avulla	- XAI:n soveltuvuus sivukanavahyökkäyksissä - LIME:n soveltuvuus suuriin malleihin ja rajoitukset - LIME:n hitaus
- Kalasteluviestien priorisointi	- Mallin tulkittavuuden hyödyt - Selitettävän mallin suorituskyky - Mallin luotettavuus - Analyytikoiden rohkaisu malleihin

TAULUKKO 11 Redusoinnin tulokset, osa 2.

- LIME:n käytön hitaus	- CTI:n merkitys kyberturvallisuuden tulevaisuudessa
- XAI-menetelmät tulevissa töissä ja tutkimuksessa	- Tekoälyn tuoma hyöty päätöksenteon nopeuteen ja turvallisuuteen - Tekoälyn tutkimuksen puute APT-kentällä
- XAI:n luoma lisäarvo CTI analyytikoille - Luottamuksen lisääntyminen	- XAI-menetelmän havainnointi ja uhkatiedon tuottamiskyky - Selittävän uhkatiedustelun hyödyt
- XAI:n hyödyntäminen pimen verkon analysoinnissa - XAI:n tuomat hyödyt nykyisille pimeän verkon CTI alustoille	- AI mallin ymmärettävyyden vaikutus proaktiiviseen oppimiskykyyn ja havainnointiin
- XAI menetelmien integrointi CTI alustoihin - Ymmärryksen lisääminen	- Datan suuri määrä - Laskentatehon tarve
- Selitysten puuttuminen tekoälystä	- Selittävän mallin tarjoama läpinäkyvyys ja luotettavuus
- XAI mallin hyödyntäminen verkon reunalla - APT:n tunnistus	- Verkon reunalla käytetyn AI:n tarjoamat edut - Verkon ja kaistan kuormituksen vähentäminen
- XAI:n hyödyt suojaustason ja puolustuskyvyn suhteen	- Nykyisten APT-havaitsemismallien haasteet - Proaktiivisuuden puute perinteisissä malleissa
- CTI:n merkitys proaktiivisena toimena - Puolustusstrategian kehittäminen	- APT-hyökkäysten puolustushaasteet kohdentamattomilla menetelmillä - Reunadatan tarjoma hyöyty uhkatiedustelulle
- Uhkatiedon tärkeys	- AI-menetelmien erkitys APT:n havaitsemisessa - Mustan laation ongelmat - Selittämättömyyden harhaanjohtaminen
- Uhkatiedon manuaalinen käsittely - AI:n potentiaali suurien tietomäärien käsittelyssä - Läpinäkyvyyden puute AI:ssa - Black-box altis harhaanjohtamiselle - Black-box altis hyökkäyksille	- Selittävä algoritmi havaintotarkkuuden opimoimiseksi - LIME:n käyttö
- CTI:n vaatimukset - XAI:n tuottamat ennustukset ja selitykset - XAI:n apu CTI:n analysoinnissa	- XAI-mallin tarjoamat hyödyt APT:n havaitsemisessa
- XAI:n vaikutus tarkkuuteen - Automaattisen uhkatideon tuottaminen	- Selitysten tarjoaminen ja luotettavuuden lisääminen
- Puolustusbudjetin ja havaitsemissuorituskyvyn välinen suhde	- Reunapuolustusmallin tarve selitettävissä olevalle uhkatiedolle
- Tasapaino puolustuksen hyötyjen ja kustannusten välillä - Strategisen poikkeaman välttäminen uhkatiedon avulla - Strategian perusta	- Black-box mallien hyvä suorituskyky ja selittämättömyys
- Selitettävän järjestelmän ennustamisen ja uhkatiedon tuottamisen kyky - Turvallisuuden kehittyminen selitettävyyden avulla - Tutkimuksen puute selitettävien mallien osalta	- Puolustusresurssien jakamisen tärkeys - Selitettävyyden kytkeytyminen analyytikoiden luottamukseen
- XAI-mallin tarjoama ennakoiva ja kestävä puolustus, tatkkuus ja selitettävyys	- Manuaalinen työ ja sen heikkous
- XAI:n tarjoamat selitykset hyökkäyksien määrittämistä ja ennusamista varten - Jatkuva oppiminen ja itsenäinen ennakoiva puolustus APT:ta vastaan	- Uhkatiedustelun selitettävyys - Uskottavuuden lisääminen
- Tarkkuuden kustannuksella vähentynyt päätösten heikkolaatuisuus	- Paikalliset yksityiset tiedot
- XAI:n määritelmä - XAI:n kyky havaita bias - Yksittäisen ennusteen selittäminen	- Havaitsemisementelmien rajoitukset - Selitettävyyden rajallisuus
- LIME:n soveltuvuus menetelmänä tekoälymallien selittämiseen	

TAULUKKO 12 Redusoinnin tulokset, osa 3.

- Selitettävissä olevan uhkatiedon merkitys	- Tekoälyn potenitaali kyberturvallisuuden ratkaisuissa - Innovointi tekoälyn ympärillä kyberturvassa
- Tulkittavuuden merkitys luotettavalle ja toimintakelpoiselle päätöksenteolle - Tulkittavuuden rooli havaitsemisjärjestelmässä	- Tarve ja mahdollisuudet tekoälylle puolustuksessa - Tekoälyn käyttämisen vaatimukset ja ylläpito
- DL-mallien suorituskyvyn ja selitettävyyden suhde	- Ihmisen roolin tärkeys tekoälyn rinnalla
- Selitettävyyden puutoksen kytkös luotettavuuteen ja uskottavuuteen	- Tekoälyn tuoma hyöty raportoinnin ja viestinnän ymmärrettävyydessä
- DL-mallin päätöksenteon tulkittavuus	- Tekoälymallien selitysten merkitys ylläpiyäjilleen
- Huomiomekanismimallin hyödyntäminen eri tehtävissä	- Kuluttajien vähäinen kiinnostus tieturvaratkaisujen selitettävyyteen - Tietoturvaratkaisun toiminnan tärkeys
- Nykyisten menetelmien puutteellinen tulkittavuus ja rajoittuneisuus	- Selitettävyyden tarve vähäpätöisemmissä vs vakavissa tilanteissa
- Tulkittava DL-malli, joka olisi luotettava	- Kuluttajien vähäinen kiinnostus tieturvaratkaisujen selitettävyyteen
- Selitettävyyden tuomat hyödyt internetin käyttäjille - Tulkittavuuden vaikutus käyttäjien valistukseen - Huijatuksi tulemisen väheneminen	- Black boxit ongelmattomuus itsessään - Selitettävyyden tarve tilanteissa, joissa malli ei toimi halutusti - Selitettävyyden tärkeys ja tarve suosittelualgoritmeissa - Ihmismielen hakkerointi
- Uhkien käsittely selitettävien oivallusten perusteella	- Somevideoiden algoritmien tarjoaman selitettävyyden tärkeys - Algoritmien ymmärtäminen ja tulkittavuus - Yhteiskunnalliset ongelmat
- Tulkittavuuden tarjoama tuki päätöksenteolle	- Selitettävyyden merkityksen korostuminen, kun kyseessä kriittinen infra tai ihmisen henki/terveys
- Suorituskyvyn ja selitettävyyden suhde - Selitettävän mallin tarvi phishingin tunnistamisessa	- Somevideoiden ja -sisällön suosittelualgoritmien selitettävyyden roolin kriittisyys - EU:n regulaation tärkeys
- Tekoälyn käyttöönotto organisaatioissa - Tekoälyn hyödyntämisen uutuus	- Sosiaalisen median tekoälyyn pohjautuvuen suosittelualgoritmien aiheuttama uhka yhteiskunnille ja demonkratialle
- Organisaatioiden IT-infrastruktuurien monimuotoisuus - Toimintaympäristön monimutkaisuus nykypäivänä	- Teknologisen kehityksen varjopuolet
- Uhkatoimijoiden ensimmäisen pääsyn menetlmien muutos vähemmän tekniikseksi	- Loppukäyttäjien kiinnostuksen puute selitettävyyttä kohtaan
- Hyökkäyksistä toipumisen suunnittelu organisaatioissa	- Päätöksissä nojautuminen pelkkään tekoälyyn - Ihmisen roolin tärkeys päätöksenteossa
- Tietoturvan merkitys organisaatiolle - Tietoturvaan suhtautuminen "pakollisena pahana" tai välttämättömyytenä - Luottamuksen merkitys tietoturvaratkaisuja kohtaan	- Selitettävyyden tärkeys ihmisen tekemän työn laadun näkökulmasta - Tekoälyn tekemien päätösten ymmärrettävyys
- Tietoturvaratkaisujen riippuvuus organisaatiosta ja sen tarpeista	- Tekoälyn rooli päätöksenteossa
- Uhkatoimijoiden hyödyntämä tekoäly - Tekoälyn hyödyntäminen hyökkäyksissä - Valmiiden työkalujen käyttö	- Selitettävyyden aiheuttamat kustannukset - Selitettävyyden merkityksen kasvu suhteessa ihmishenkeen
- Tekoälyn hyödyntäminen lähitulevaisuudessa hyökkäyksissä	

TAULUKKO 13 Redusoinnin tulokset, osa 4.

- Black boxiin nojaavien mallien epäonnistuminen - XAI-mallien suosion kasvaminen	- ML-mallien selittämättömyys
- Tulkittavuuden tuomat hyödyt kausaalisuhteiden tunnistamiseen	- Selitysten merkitys korkean riskin päätöksissä - Järjestelmän tehokkuuden kyseenalaisuus ilman selityksiä
- ML-järjestelmien luottamuksen ja oikeudenmukaisuuden varmistaminen - Kestävyyden ja luotettavuuden kriittisyys korkean riskin ympäristöissä - Koneiden tekemien päätösten seurausten merkitys	- XAI:n tarjoama vastaus ML-mallien selittämättömyyteen
- Perinteisten ML-mallien selittämättömyyden aiheuttama hämmenys käyttäjälleen - XAI-mallien kehittynyt menettely	- XAI:n merkityksellisyys tulosten hallinnassa - Perustelut XAI:n luotettavuudelle
- Selitysten antama syy ennusteelle	- Selitettävyyden ohittaminen - XAI:n rooli kriittisessä päätöksenteossa
- ML mallien selittämisen moninaiset menetelmät - LIME	- ML:n suuri merkitys kyberturvallisuuden ratkaisuissa nykypäivänä
- Black box lähestymistavan aiheuttaman epämukavuuden laajuus	- Black box mallien käyttö - Selitettävyyden varmistaminen
- XAI:n lisäämä tekoälymallin tarkkuus	- Kyberhyökkäykset selitettäviä avoimia malleja kohtaan
- Luotettamuksen ja läpinäkyvyyden parantaminen selitysten avulla	- XAI:n kohtaamat kyberuhat
- XAI-mallien potentiaali havaitsemisen paratamisessa - Ennustustulosten luotettavuus - Selitettävyyden tärkeys kriittisen informaation käsittelyssä	- XAI:n kohtaamien kyberuhkien torjuminen
- Selitettävyyden tarjoama hyöty OSINT:ssa - Ennakointi	- Selitysmenetelmien arvionnin tärkeys
- Nykyisten kyberturvallisuuden ML-mallien puutteet - Selitettävyyden suuri tärkeys	- XAI-mallien testaaminen ihmisellä - Vaikutus päätöksentekoon
- XAI-mallien hyvä saatavuus tänä päivänä	- Tulkittavuuden ja selitettävyyden erot
- XAI-mallien kyky paljastaa ML-mallien tapa toimia	- XAI-mallien manipuolointi - XAI-mallien luotettavuus - Selitysten herkkyys
- Selitettävyysmenetelmien haasteet - Eksplisiittisen esityksen merkitys	- Hyökkäystyypit XAI-malleja kohtaan
- XAI-menetelmien hankala tulkittavuus ihmiselle - Oppimismenetelmien rooli kehitys - Selitysten tarve päätöksentekojärjestelmissä	- Selitettävien menetelmien turvallisuus kyberturvan alalla
- AI-järjestelmien suunnittelun perimmäinen tavoite muualla kuin selityksissä	- Kontrafaktuaalisten selitysten hyödyntäminen hyökkäyksissä
- Ihmisen ja koneen välisen vuorovaikutuksen merkitys	- Koulutusdataan kohdistuva uhka
- AI:n kehityksen riippuvuus ML:stä	- Selitysmallien varastaminen
- Kyberturvallisuusratkaisuissa käytettyjen black box menetelmien uskottavuuden vähyys - XAI:n roolin merkitys älykkäissä kaupungeissa	- Selitysmallin päättelyyn liittyvät hyökkäykset
- Kyberturvallisuusalan jännite turvallisuuden ja käytettävyyden välillä - Sen ilmeneminen myös selitysten alalla - Selitysten tarjoama apu sisäisten ominaisuuksien tutkimiseen - Hyökkäyspinnan laajeneminen selitettävyyden myötä	- Myrkytyshyökkäykset - Harjoitusdatan saastuttaminen
- XAI-melleihin kohdistuvat hyökkäykset - Kasvojentunnistus	

TAULUKKO 14 Redusoinnin tulokset, osa 5.

- Pelkkiin koneoppimismalleihin luottamisen hatara pohja - Tekoälyn tekemiin ratkaisuihin nojaaminen oikeuden edessä	- XAI-mallin suorituskykyisyys
- Tekoälymallien kehittymisen aste - Generatiivisten tekoälymallien epäluotettavuus - Ihmisen rooli todistamisessa	- XAI-mallin adaptoituminen uhkaan - OSINT-tiedolla rikastaminen
- Selitettävyyden kysynnän kasvu suhteessa käytettävän kohteen kriittisyyteen - Kiinnostuksen taso selitettävyyttä kohtaan, jos malli vain toimii ja tekee tehtävänsä	- XAI-mallin oppiminen ja mukautuminen - XAI-mallin keskeneräisyys
- Ihmisten liika luottamus koneoppimista kohtaan	- OSINT:in hyödyntäminen syväoppimisanalyysissa - XAI-mallin haitat - Selityksien puute ja haaste
- Selitettävyyden tarve	- Tulevaisuuden tutkimuksessa tarvitaan lisää OSINT-tietoa - Selitysten kehittäminen - Selitysten vajavaisuus
- Tekoälyn vinoumat - Selittämättömyyden puutteen merkitys harhaoihin - Puolueellinen päätöksenteko	- Tekoälyjärjestelmien selitysten tarve - Selitysten välttämättömyys tietyillä aloilla
- Tekoälyn vinoumien ja puolueellisuuden vaikutukset tiettyihin ihmisryhmiin	- Läpinäkyvyyden vaikutus ihmishenkiin
- XAI:n haavoittuvuudet ja selitysmallien manipuloimisen mahdollisuus	- XAI:n tuoma ratkaisu black boxin ongelmaan
- XAI-malleihin luottaminen - XAI:n hyöty, jos se on haavoittuva ja sitä voidaan manipuloida	- XAI:n mullistava vaikutus ML-tekniikoihin - Ennustustarkkuuden säilyminen - Läpinäkyvyyden vaikutus luotettavuuteen ja ymmärrykseen
- Yliluottaminen XAI-malleihin ja sen riskit	- XAI:n tuoma apu järjestelmäsuunnittelijoille - Jälkikäteen annettujen selitysten hyödyt
- Nykyinenkin tietoturva heikkoa - Koneoppimismallien suojaaminen	- ML-perustaisten IDS-järjestelmien kehittyminen - Hyökkääjien oveluus ohittaa nykyiset mallit - XAI:n tuoma hyöty proaktiiviseen toimintaan
- Koneoppimismallien rooli tärkeissä päätöksissä liittyen ihmisten elämiin - Muiden tieteenalojen yhdistäminen tekoälyn kehitykseen kuin insinööritieteiden	- IDS-järjestelmien epäonnistuminen - False positiven fataalius - Selitysten suuri tarve ja niiden välttämättömyys
- Tekoälyn hyödyntäminen kyberuhkien analysoinnissa - Nykyisten menetelmät black boxeja - Selittämättömyyden aiheuttama epävarmuus	- Selitysten tuoma hyöty IDS-järjestelmiin
- Kysymykset liittyen tietoturvassa käytettyjen tekoälymallien selittämättömyyteen	- Selitysten merkittävä rooli verkkohyökkäysten ennustamisessa
- Eri tieteenaloilla käytetty huomiomekanismi - Neuroverkkomallien tulkinta	- XAI:n tuomat hyödyt IDS-järjestelmissä - Selitettävyyden lisääntyminen CIA:n kanssa
- Huomiomekanismin selitettävyys - Ihmisenkaltainen selitettävyys	- XAI:n tehokkuus hyökkäyksien tunnistamisessa - Shapley-arvojen käyttö XAI-mallissa - Ymmärrys, miksi jokin malli päätyy johonkin ennusteeseen
- OSINT:n ja CTI:n hyödyntäminen proaktiivisena toimena	- Shapley-arvojen tuoma hyöty tietoverkkoriskien selvittämiseen
- OSINT ja CTI raporttien määrän kasvu - Neuroverkkomallin hyödyntäminen haittaohjelmien analyysissa	- XAI:n tuoma hyöty haitallisen liikenteen tunnistamiseen
- MITRE ATT&CK -kehyksen vaikutus syväoppimisen tulosten tulkittavuuteen	- ML mallien mustat laatikot - Läpinäkymättömyyden aiheuttamat riskit
- MITRE ATT&CK:n rajoitukset - Pelkästään ATT&CK:iin nojaamisen vaaranpaikat - Suorituskyvyn parantaminen muilla OSINT-tiedoilla ja lähteillä	- ML:lle syötetyn datan muuttumisen aiheuttamat ongelmat

TAULUKKO 15 Klusteroinnin tulokset osa 1.

Pelkistetyt ilmaukset	Alaluokka
Käytetyn havaitsemismallin luottamuksen kasvattaminen Selitettävän havainnointimallin luotettavuus Yleistettävyys Mallin luotettavuus Selitettävyyden varmistaminen Perustelut XAI:n luotettavuudelle Selitettävyyden tärkeys kriittisen informaation käsittelyssä Ennustustulosten luotettavuus Luotettamuksen ja läpinäkyvyyden parantaminen selitysten avulla Analyytikoiden rohkaisu malleihin Selitettävyyden merkityksen korostuminen, kun kyseessä kriittinen infra tai ihmisen henki/terveys Tietoturvaratkaisun toiminnan tärkeys Luottamuksen merkitys tietoturvaratkaisuja kohtaan Tulkittavuuden vaikutus käyttäjien valistukseen Selitettävyyden puutoksen kytkös luotettavuuteen ja uskottavuuteen DL-mallin päätöksenteon tulkittavuus Uskottavuuden lisääminen Selitettävyyden kytkeytyminen analyytikoiden luottamukseen Selitysten tarjoaminen ja luotettavuuden lisääminen AI-menetelmien merkitys APT:n havaitsemisessa Ymmärrys, miksi jokin malli päätyy johonkin ennusteeseen Selittävän mallin tarjoama läpinäkyvyys ja luotettavuus Selitettävyyden vaikutus mallin luotettavuuteen ML-järjestelmien luottamuksen ja oikeudenmukaisuuden varmistaminen Luottamuksen lisääntyminen XAI-mallien suosion kasvaminen Turvallisuuden kehittyminen selitettävyyden avulla Puolustusstrategian kehittäminen AI mallin ymmärettävyyden vaikutus proaktiiviseen oppimiskykyyn ja havainnointiin	Luottamuksen vahvistaminen
Johdonmukainen suorituskyky luokittelussa Selitettävän havainnointimallin suorituskyky Selitettävän havainnointimallin tehokkuus Suorituskyvyn ja selitettävyyden suhde XAI-mallien potentiaali havaitsemisen parantamisessa Kestävyyden ja luotettavuuden kriittisyys korkean riskin ympäristöissä ML-perustaisten IDS-järjestelmien kehittyminen Selitettävän mallin suorituskyky XAI käsitteen historia XAI:n vaikutus tarkkuuteen Automaattisen uhkatideon tuottaminen XAI sovellusten käyttö yleisesti XAI menetelmien jakautuminen kahteen kategoriaan LIME:en pohjautuvan XAI-puolustustekniikan tehokkuus XAI menetelmien integrointi CTI alustoihin CTI:n vaatimukset XAI:n tuottamat ennustukset ja selitykset XAI:n apu CTI:n analysoinnissa XAI-mallin tarjoama ennakoiva ja kestävä puolustus, tarkkuus ja selitettävyys Jatkuva oppiminen ja itsenäinen ennakoiva puolustus APT:ta vastaan XAI:n kyky havaita bias Yksittäisen ennusteen selittäminen XAI-mallien kehittynyt menettely XAI:n tehokkuus hyökkäyksien tunnistamisessa XAI-mallin suorituskykyisyys Shapley-arvojen tuoma hyöty tietoverkkoriskien selvittämiseen	Tehokkuus

TAULUKKO 16 Klusteroinnin tulokset, osa 2.

Tekoälyn tuoma hyöty päätöksenteon nopeuteen ja turvallisuuteen Tulkittavuuden merkitys luotettavalle ja toimintakelpoiselle päätöksenteolle Tulkittavuuden tarjoama tuki päätöksenteolle Tekoälyn tuoma hyöty raportoinnin ja viestinnän ymmärrettävyydessä XAI:n rooli kriittisessä päätöksenteossa Selitysten merkitys korkean riskin päätöksissä Selitysten tarve päätöksentekojärjestelmissä Ennakointi Läpinäkyvyyden vaikutus luotettavuuteen ja ymmärrykseen Tekoälymallien selitysten merkitys ylläpitäjilleen Ihmisen roolin tärkeys päätöksenteossa Selitysten antama syy ennusteelle Päätöksissä nojautuminen pelkkään tekoälyyn	Päätöksenteon tukeminen
XAI menetelmät piirteiden havaitsemisessa XAI menetelmien arviointi XAI:n hyödyntäminen pimeän verkon analysoinnissa XAI:n lisäämä tekoälymallin tarkkuus Selitettävyyden ohittaminen XAI-mallien hyvä saatavuus tänä päivänä XAI mallin hyödyntäminen verkon reunalla ML mallien selittämisen moninaiset menetelmät XAI-mallin oppiminen ja mukautuminen Selitettävyyden merkityksen kasvu suhteessa ihmishenkeen APT:n tunnistus Somevideoiden ja -sisällön suosittelualgoritmien selitettävyyden roolin kriittisyys LIME:n soveltuvuus menetelmänä tekoälymallien selittämiseen Shapley-arvojen käyttö XAI-mallissa XAI-menetelmän havainnointi ja uhkatiedon tuottamiskyky Selitettävyyden tarve vähäpätöisemmissä vs vakavissa tilanteissa LIME:n käyttö	XAI-menetelmien moninaisuus

TAULUKKO 17 Klusteroinnin tulokset, osa 3.

Suostutteluvihjeet selitettävyyttä tukemassa Käytetyn havaitsemismallin adaptoituminen ihmismäiseen käytökseen Parempi havainnointikyky Mallin tulkittavuuden hyödyt XAI:n tarjoama vastaus ML-mallien selittämättömyyteen Selitysten tarjoama apu sisäisten ominaisuuksien tutkimiseen XAI-mallien kyky paljastaa ML-mallien tapa toimia Tulkittavuuden tuomat hyödyt kausaalisuhteiden tunnistamiseen Selitysten tuoma hyöty IDS-järjestelmiin XAI:n tuoma hyöty proaktiiviseen toimintaan Jälkikäteen annettujen selitysten hyödyt XAI-mallin adaptoituminen uhkaan Neuroverkkomallin hyödyntäminen haittaohjelmien analyysissa XAI:n hyödyntäminen identiteetin ja yksityisyyden suojaamisessa XAI-mallin tarjoamat hyödyt APT:n havaitsemisessa XAI:n tuomat hyödyt nykyisille pimeän verkon CTI alustoille XAI:n käyttö puolustusstrategioissa ROAR-metriikan käyttö ja ominaisuudet Huijatuksi tulemisen väheneminen XAI:n tuoma hyöty haitallisen liikenteen tunnistamiseen ROAR selitettävien menetelmien arvioimisessa ROAR:n XAI menetelmien tehokkuuden arvioinnissa Ehdotetut XAI-algoritmit XAI:n tuomat hyödyt IDS-järjestelmissä XAI:n luoma lisäarvo CTI analyytikoille Ennustustarkkuuden säilyminen Algoritmien soveltuvuus erilaisiin malleihin LIME ja Saliency Map -mallien soveltaminen WF-hyökkäyksein tutkimiseen Selitettävyyden tuomat hyödyt internetin käyttäjille Mallin käyttäytymisen selitettävyys Mallin tulkinnan helppous XAI:n tuoma apu järjestelmäsuunnittelijoille LIME:n käyttö selittämisessa Ymmärryksen lisääminen XAI:n hyödyt suojaustason ja puolustuskyvyn suhteen XAI:n tarjoamat selitykset hyökkäyksien määrittämistä ja ennusamista varten Tulkittavuuden rooli havaitsemisjärjestelmässä	XAI-mallien ja menetelmien hyödyt
Havaitsemisementelmien rajoitukset Selitettävyyden rajallisuus Selitettävyysmenetelmien haasteet	Selitysten rajoitukset
RF-mallin väärinluokitteluaste XAI:n soveltuvuus sivukanavahyökkäyksissä LIME:n soveltuvuus suuriin malleihin ja rajoitukset LIME:n hitaus Kyberturvallisuusalan jännite turvallisuuden ja käytettävyyden välillä Sen ilmeneminen myös selitysten alalla Selitysten vajavaisuus XAI-mallin haitat XAI-mallin keskeneräisyys Yliluottaminen XAI-malleihin ja sen riskit XAI:n hyöty, jos se on haavoittuva ja sitä voidaan manipuloida XAI-malleihin luottaminen XAI:n haavoittuvuudet ja selitysmallien manipuloimisen mahdollisuus Selitettävyyden aiheuttamat kustannukset LIME:n käytön hitaus Kuluttajien vähäinen kiinnostus tieturvaratkaisujen selitettävyyteen Nykyinenkin tietoturva heikkoa Nykyisten APT-havaitsemismallien haasteet Laskentatehon tarve Kuluttajien vähäinen kiinnostus tieturvaratkaisujen selitettävyyteen Datan suuri määrä Loppukäyttäjien kiinnostuksen puute selitettävyyttä kohtaan	XAI-mallien haasteet

TAULUKKO 18 Klusteroinnin tulokset, osa 4.

Pelkkiin koneoppimismalleihin luottamisen hatara pohja Tekoälyn tekemiin ratkaisuihin nojaaminen oikeuden edessä Kiinnostuksen taso selitettävyyttä kohtaan, jos malli vain toimii ja tekee tehtävänsä Koneoppimismallien rooli tärkeissä päätöksissä liittyen ihmisten elämiin XAI:n merkityksellisyys tulosten hallinnassa XAI:n roolin merkitys älykkäissä kaupungeissa AI-järjestelmien suunnittelun perimmäinen tavoite muualla kuin selityksissä Selitettävyyden lisääntyminen CIA:n kanssa	Tekoälyn rooli
Selitettävyyden tärkeys ihmisen tekemän työn laadun näkökulmasta Tekoälyn tekemien päätösten ymmärrettävyys Tekoälyn rooli päätöksenteossa Ihmisen rooli todistamisessa Vaikutus päätöksentekoon Ihmisen ja koneen välisen vuorovaikutuksen merkitys Opiimismenetelmien roolin kehitys Ihmisenkaltainen selitettävyys Huomiomekanismin selitettävyys XAI-mallien testaaminen ihmisellä Puolueellinen päätöksenteko Tekoälyn vinoumien ja puolueellisuuden vaikutukset tiettyihin ihmisryhmiin Ihmisten liika luottamus koneoppimista kohtaan	Tekoälyn ja ihmisen välinen vuorovaikutus
Suostutteluvihjeiden käyttö phishing-sähköpostien havaitsemisessa Tarve selitettäville lähestymistavoille Selitettävyyden hyödyt phishingiin liittyvässä kyberuhkatiedustelussa Selitettävyys lisää ymmärrystä mallin tuottamista ratkaisuista Nykyisten kyberturvallisuuden ML-mallien puutteet Selitysten suuri tarve ja niiden välttämättömyys IDS-järjestelmien epäonnistuminen Selitysten välttämättömyys tietyillä aloilla Tekoälyjärjestelmien selitysten tarve Selitettävyyden tarve Selitettävyyden suuri tärkeys Selitettävyyden kysynnän kasvu suhteessa käytettävän kohteen kriittisyyteen Proaktiivisuuden puute perinteisissä malleissa Selitettävän mallin tarve phishingin tunnistamisessa Koneiden tekemien päätösten seurausten merkitys XAI:n tuoma ratkaisu black boxin ongelmaan Selitettävän mallin toiminta verrattuna perinteisiin luokittelumalleihin Proaktiivisuuden puute perinteisissä malleissa ML mallien käsitteleminen useimmiten selitettävinä malleina Kalasteluviestien priorisointi Selittävä algoritmi havaintotarkkuuden opimoimiseksi Algoritmien ymmärtäminen ja tulkittavuus Suositus organisaatioiden strategioiden kehittämiseksi selitettävän AI mallin avulla Mallin tuottaman tuloksen selitettävyys Selitettävyyden tarve tilanteissa, joissa malli ei toimi halutusti Nykyisten menetelmien puutteellinen tulkittavuus ja rajoittuneisuus XAI:n mullistava vaikutus ML-tekniikoihin XAI kyberturvallisuuden alalla alustavasti Black-box altis harhaanjohtamiselle Black-box altis hyökkäyksille Somevideoiden algoritmien tarjoaman selitettävyyden tärkeys Selitettävyyden tärkeys ja tarve suosittelualgoritmeissa	XAI mallien tarve

TAULUKKO 19 Klusteroinnin tulokset, osa 5.

Strategisen poikkeaman välttäminen uhkatiedon avulla Strategian perusta CTI:n merkitys proaktiivisena toimena Tulkittava DL-malli, joka olisi luotettava Selitettävyyden tarjoama hyöty OSINT:ssa Läpinäkyvyyden vaikutus ihmishenkiin OSINT:in hyödyntäminen syväoppimisanalyysissa OSINT-tiedolla rikastaminen MITRE ATT&CK:n rajoitukset Pelkästään ATT&CK:iin nojaamisen vaaranpaikat OSINT ja CTI raporttien määrän kasvu OSINT:n ja CTI:n hyödyntäminen proaktiivisena toimena Reunadatan tarjoma hyöyty uhkatiedustelulle Yhteiskunnalliset ongelmat Uhkien käsittely selitettävien oivallusten perusteella Reunapuolustusmallin tarve selitettävissä olevalle uhkatiedolle Uhkatiedustelun selitettävyys Suorituskyvyn parantaminen muilla OSINT-tiedoilla ja lähteillä Selitettävän järjestelmän ennustamisen ja uhkatiedon tuottamisen kyky MITRE ATT&CK -kehyksen vaikutus syväoppimisen tulosten tulkittavuuteen CTI:n merkitys kyberturvallisuuden tulevaisuudessa Selittävän uhkatiedustelun hyödyt Selitettävissä olevan uhkatiedon merkitys	XAI:n ja kynberuhkatideustelun yhdistäminen
Hyökkäyksistä toipumisen suunnittelu organisaatioissa Tietoturvan merkitys organisaatiolle Tietoturvaan suhtautuminen ""pakollisena pahana"" tai välttämättömyytenä Tietoturvaratkaisujen riippuvuus organisaatiosta ja sen tarpeista Ihmismielen hakkerointi Teknologisen kehityksen varjopuolet Sosiaalisen median tekoälyyn pohjautuvuen suosittelualgoritmien aiheuttama uhka yhteiskunnille ja demonkratialle EU:n regulaation tärkeys Tekoälyn hyödyntäminen kyberuhkien analysoinnissa	Toimintaympäristön muutos
Uhkatoimijoiden hyödyntämä tekoäly Tekoälyn hyödyntäminen hyökkäyksissä Valmiiden työkalujen käyttö Tekoälyn hyödyntäminen lähitulevaisuudessa hyökkäyksissä Hyökkääjien oveluus ohittaa nykyiset mallit	Hyökkääjien käyttämä tekoäly

TAULUKKO 20 Klusteroinnin tulokset, osa 6.

Uhkatoimijoiden ensimmäisen pääsyn menetlmien muutos vähemmän tekniikseksi	Muuttuvat uhat ja hyökkäysvektorit
Organisaatioiden IT-infrastruktuurien monimuotoisuus Toimintaympäristön monimutkaisuus nykypäivänä	Tekoälyn implementoinnin haasteet
Manuaalisen ihimistyön kuormittavuus Perinteisten ML-mallien vaatima manuaalinen työ Vanhat mallit raskaita Uhkatiedon tärkeys ML:n suuri merkitys kyberturvallisuuden ratkaisuissa nykypäivänä Uhkatiedon manuaalinen käsittely Tekoälyn potenitaali kyberturvallisuuden ratkaisuissa Tekoälyn käyttöönotto organisaatioissa Manuaalinen työ ja sen heikkous Verkon reunalla käytetyn AI:n tarjoamat edut Tarve ja mahdollisuudet tekoälylle puolustuksessa Selitysten merkittävä rooli verkkohyökkäysten ennustamisessa AI:n potentiaali suurien tietomäärien käsittelyssä	Tekoälymallien tarve kyberpuolustuskessa
Vastatoimien kehittäminen Ymmärrys hyökkääjän taktiikoista Teoreettisen näkökulman tuoma hyöty selitettävien mallien tuottamisessa Black-box-mallien tulkinta Selitettävien menetelmien turvallisuus kyberturvan alalla Selitysmenetelmien arvionnin tärkeys AI:n kehityksen riippuvuus ML:stä XAI-menetelmien hankala tulkittavuus ihmiselle Eksplisiittisen esityksen merkitys Selityksien puute ja haaste Neuroverkkomallien tulkinta Eri tieteenaloilla käytetty huomiomekanismi Tekoälyn vinoumat Aikaisemman tutkimuksen puute Ihmisen roolin tärkeys tekoälyn rinnalla Innovointi tekoälyn ympärillä kyberturvassa Tekoälyn hyödyntämisen uutuus DL-mallien suorituskyvyn ja selitettävyyden suhde Tekoälyn tutkimuksen puute APT-kentällä Poikkitieteellisen tutkimuksen tarve Selittämisen ymmärtäminen Tutkimuksen puute selitettävyyden osalta XAI mallien tutkimuksen puute WF-hyökkäyksissä Tulevaisuuden tutkimuksessa tarvitaan lisää OSINT-tietoa XAI käsitteen historia Selitysten kehittäminen Miten XAI:ta tutkittu kyberturvallisuuden alalla tähän mennessä XAI:n tarkoitus mallien käytöksen tutkimisessa XAI-menetelmät tulevissa töissä ja tutkimuksessa Puolustusbudjetin ja havaitsemissuorituskyvyn välinen suhde Tasapaino puolustuksen hyötyjen ja kustannusten välillä Tutkimuksen puute selitettävien mallien osalta	Tutkimuksen tarve

TAULUKKO 21 Klusteroinnin tulokset, osa 7.

Ristiriita, mitä selitettävyys on tieteellisen ymmärryksen näkökulmasta XAI analyysin haasteet Puolustusresurssien jakamisen tärkeys Tekoälyn käyttämisen vaatimukset ja ylläpito Tulkittavuuden ja selitettävyyden erot False positiven fataalius Muiden tieteenalojen yhdistäminen tekoälyn kehitykseen kuin insinööritieteiden Koneoppimismallien suojaaminen Generatiivisten tekoälymallien epäluotettavuus Tekoälymallien kehittymisen aste	Tuktimushaasteet
Kyberhyökkäykset selitettäviä avoimia malleja kohtaan XAI:n kohtaamat kyberuhat XAI-mallien manipuolointi XAI-mallien luotettavuus XAI-melleihin kohdistuvat hyökkäykset Hyökkäyspinnan laajeneminen selitettävyyden myötä Myrkytyshyökkäykset Selitysmallin päättelyyn liittyvät hyökkäykset Harjoitusdatan saastuttaminen Selitysmallien varastaminen Koulutusdataan kohdistuva uhka Kontrafaktuaalisten selitysten hyödyntäminen hyökkäyksissä Selitysten herkkyys Hyökkäystyypit XAI-malleja kohtaan	XAI:n haavoittuvuudet
Aiemmin käytettyjen DL mallien vaikea ymmärrettävyys kalasetun havaitsemisessa ML ja DL -mallien vaikea selitettävyys Selitysten puuttuminen tekoälystä Tarkkuuden kustannuksella vähentynyt päätösten heikkolaatuisuus XAI:n kohtaamien kyberuhkien torjuminen Black box mallien käyttö Black box lähestymistavan aiheuttaman epämukavuuden laajuus Perinteisten ML-mallien selittämättömyyden aiheuttama hämmenys käyttäjälleen Black boxiin nojaavien mallien epäonnistuminen Black-box mallien hyvä suorituskyky ja selittämättömyys	Vaikea ymmärrettävyys
DL mallit ovat black-box malleja DL mallien heikko tulkittavuus monimumtkaisuuden vuoksi Läpinäkyvyyden puute AI:ssa Mustan laation ongelmat Järjestelmän tehokkuuden kyseenalaisuus ilman selityksiä ML-mallien selittämättömyys Kyberturvallisuusratkaisuissa käytettyjen black box menetelmien uskottavuuden vähyys ML:lle syötetyn datan muuttumisen aiheuttamat ongelmat Läpinäkymättömyyden aiheuttamat riskit ML mallien mustat laatikot Kysymykset liittyen tietoturvassa käytettyjen tekoälymallien selittämättömyyteen Selittämättömyyden aiheuttama epävarmuus Selittämättömyyden puutteen merkitys harhoihin Selittämättömyyden harhaanjohtaminen Nykyisten menetelmät black boxeja Black boxien aiheuttamat ongelmat	Läpinäkyvyyden puute
DL mallien selitettävyys edistynyt XAI:n harvinaisuus tietoturvasovelluksissa	Selitettävyyden kehitys

TAULUKKO 22 Abstrahointi, osa 1.

Alaluokka	Yläluokka
Luottamuksen vahvistaminen Tehokkuus Päätöksenteon tukeminen XAI-mallien ja menetelmien hyödyt XAI:n ja kynberuhkatideustelun yhdistäminen XAI-menetelmien moninaisuus	XAI:n tuomat edut ja mahdollisuudet kyberturvallisuuden kentälle
Selitysten rajoitukset XAI-mallien haasteet XAI:n haavoittuvuudet Tekoälyn implementoinnin haasteet Hyökkääjien käyttämä tekoäly	XAI-mallien keskeneräisyys ja siihen liittyvät haasteet
Vaikea ymmärrettävyys Läpinäkyvyyden puute	Perinteisten AI/ML-mallien avoimuuden ja tulkittavuuden puuttuminen
Toimintaympäristön muutos Muuttuvat uhat ja hyökkäysvektorit Selitettävyyden kehitys	Jatkuvasti toimintaympäristö ja kyberturvallisuuskentän muutostila ja siihen vastaamisen tarve
XAI mallien tarve Tekoälyn rooli Tekoälyn ja ihmisen välinen vuorovaikutus	XAI:n paikka ja rooli kyberturvallisuuden ratkaisuissa
Tekoälymallien tarve kyberpuolustuskessa Tutkimuksen tarve Tuktimushaasteet	Tutkimuksen puute ja siihen liittyvät haasteet

TAULUKKO 23 Abstrahointi, osa 2.

Alaluokka	Yläluokka	Yhdistävä luokka
Luottamuksen vahvistaminen Tehokkuus Päätöksenteon tukeminen XAI-mallien ja menetelmien hyödyt XAI:n ja kynberuhkatideustelun yhdistäminen XAI-menetelmien moninaisuus	XAI:n tuomat edut ja mahdollisuudet kyberturvallisuuden kentälle	**XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa**
XAI mallien tarve Tekoälyn rooli Tekoälyn ja ihmisen välinen vuorovaikutus	XAI:n paikka ja rooli kyberturvallisuuden ratkaisuissa	
Vaikea ymmärrettävyys Läpinäkyvyyden puute	Perinteisten AI/ML-mallien avoimuuden ja tulkittavuuden puuttuminen	**Perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan**
Toimintaympäristön muutos Muuttuvat uhat ja hyökkäysvektorit Selitettävyyden kehitys	Jatkuvasti toimintaympäristö ja kyberturvallisuuskentän muutostila ja siihen vastaamisen tarve	
Tekoälymallien tarve kyberpuolustuskessa Tutkimuksen tarve Tuktimushaasteet	Tutkimuksen puute ja siihen liittyvät haasteet	**XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla**

6 POHDINTA JA JOHTOPÄÄTÖKSET

Tässä luvussa esitellään tutkimuksen pohdinta sekä johtopäätökset. Ensin syvennytään pohdintaan tutkimusta eettisyyden ja luotettavuuden näkökulmasta. Sen jälkeen siirrytään tarkastelemaan kirjallisuutta, johon tutkimus nojaa. Kolmanneksi tarkastellaan tutkimuksen tuloksia aiempaa tietoa sekä tämän pro gradu -tutkielman teoreettista viitekehystä vasten. Lopuksi avataan tutkimuksen johtopäätökset.

6.1 Pohdinta

Pohdinta on jaettu kolmeen osaan. Tuomi ja Sarajärvi (2018, s. 182) kuvaavat laadullisen tutkimuksen pohdinnan tämän kaltaisella jaottelulla. He kuvailevat samalla laadullisen tutkimuksen pohdintaa joustavaksi (Tuomi & Sarajärvi, 2018, s. 182).

6.1.1 Tutkimuksen luotettavuus ja eettisyys

Tuomi ja Sarajärvi (2018, s. 182) nostavat eettisyyden ja luotettavuuden tarkastelun näkökulmasta eettisen kestävyyden tärkeimmäksi seikaksi. Tällä tarkoitetaan käytännössä sitä, että mikäli tutkimus ei ole eettisesti kestävä, se ei voi myöskään olla luotettava. Samaan aikaan eettinen kestävyys ei kuitenkaan takaa sitä, että tutkimus olisi automaattisesti luotettava. Luotettavuuden kriteeristö on esitetty taulukossa 25 (Tuomi & Sarajärvi, 2018, s. 162).

Tutkimusmenetelmien luotettavuutta tarkastellaan yleensä validiteetin ja reliabiliteetin näkökulmasta. Validiteetilla tarkoitetaan, että tutkimuksessa on tutkittu sitä, mitä on luvattu tutkia. Reliabiliteetti taas kuvaa sitä, miten tutkimustulokset ovat toistettavissa. Laadullisen tutkimuksen kentällä näitä käsitteitä kohtaan on kuitenkin kohdistettu kritiikkiä. Kritisointi johtuu siitä, että kyseiset käsitteet ovat saaneet alkunsa määrällisen tutkimuksen alueella ja ne vastaavat määrällisen tutkimuksen asettamia tarpeita. Onkin esitetty monissa laadullista tutkimusta käsittelevissä oppaissa, että nämä käsitteet joko hylätään tai

vaihtoehtoisesti korvataan, kun tehdään laadullisen tutkimuksen arvioimista. Taulukko 25 pyrkii vastaamaan tähän haasteeseen. (Tuomi & Sarajärvi, 2018, s. 160–161.)

Seuraavaksi tarkastellaan tutkimusta taulukkoa 25 vasten alkaen uskottavuudesta ja vastaavuudesta ylhäältä alas. Uskottavuutta tarkasteltaessa voidaan tehdä toteamus, että luotettavuuden yksinä osatekijöinä olevat sovellettavuus ja pysyvyys asettuvat ainakin osittain kriittiseen valoon tämän tutkielman tapauksessa. Juuri tätä nimenomaista tutkimusta voi olla hankalaa soveltaa ja toistaa sellaisenaan jotain muuta aihetta tutkiessa. Tutkimusaineisto on hyvin spesifi ja melko kapealla kärjellä valikoitu, joten toista aihetta vasten tutkimuksen toistaminen voisi olla haastavaa. Jotta tutkimus olisi toistettavampi, niin esimerkiksi huomattavasti laajempi kysely- tai haastattelututkimus voisi olla tässä tapauksessa toistettavampi.

Pysyvyyden näkökulmasta tutkimus tullee kestämään aikaa, vaikka teknologiat tekoälyn ympärillä kehittyvätkin kovaa vauhtia. Tekoälyn fundamentit ja ihminen sen pääasiallisena käyttäjänä tulevat todennäköisesti pysymään tulevaisuudessakin, joten tämä tutkimus on siitä näkökulmasta tarkasteltuna relevantti vielä vuosienkin päästä.

Tutkimukseen osallistuneen haasteltavan kuvaus on tutkimuksen luotettavuuden kannalta riittävä, ja lukija saa hyvän käsityksen haastateltavasta. Kerätyn aineiston totuudenmukaisuutta arvioita pohtiessa voitaisiin todeta, että aineisto on totuudenmukaista. Kaikki kirjallinen aineisto on kerätty tunnetuista ja luotettavista tietokannoista ja lähteistä, minkä lisäksi aineistona käytetyt artikkelit ovat vertaisarvioituja. Nämä tekijät lisäävät aineiston ja sitä myötä koko tutkimuksen uskottavuutta.

Vastaavuutta tarkasteltaessa taulukossa 25 puhutaan tutkittavista, joita tämä tutkimuksen tapauksessa on vain yksi, jos asiaa katsotaan haastattelujen näkökulmasta. Tämä asettaa vastaavuuden arvioinnille haasteen, koska kerätty aineisto koostuu suurimmaksi osaksi kirjallisesta aineistosta. Tutkijan tuottamat rekonstruktiot vastaavat alkuperäisiä konstruktioita hyvin siitä näkökulmasta tarkasteltuna, että alkuperäinen data on onnistuttu säilyttämään koko aineiston analyysiprosessin ajan. On kuitenkin mahdollisuus sille, että joitain yksinkertaistuksia on tapahtunut, joka osiltaan laskee vastaavuuden tasoa. Analyysiprosessi on kuitenkin toteutettu kurinalaisesti ja polku alkuperäiseen dataan säilyttäen, joten sillä on vastaavuuden suhteen positiivinen vaikutus. Tutkimuksen tulososiossa analyysiprosessi on kuvattu läpinäkyvästi.

Tulosten siirrettävyys on tämän tutkimuksen tapauksessa haasteellista ja hieman ongelmallista. Peruste tälle pohdinnalle on, että kyseinen tutkimuskonteksti on sen verran vähän tutkittu ja spesifi, että johonkin ulkopuoliseen vastaavaan kontekstiin tulosten siirtäminen voisi olla monimutkaista. Tietyin ehdoin tulokset kuitenkin voisivat olla siirrettävissä. Siirrettävyys riippuu paljolti siitä, kuinka hyvin tutkittavat ympäristöt vastaavat toisiaan.

Luotettavuuden näkökulmasta tutkimuksen voidaan katsoa olevan vankalla pohjalla. Tämän pro gradu -tutkielman ohjaaja on tarkastanut tutkimusprosessin toteutumisen. Analyysiprosessista on käyty tutkijan ja ohjaajan välillä kommenttien ja palautteen vaihtoa useassa eri prosessin vaiheessa. Prosessin

etenemistä on myös kuvattu ohjaajalle konkreettisin esimerkein analyysin edetessä.

Tutkimuksessa on otettu huomioon ulkoista vaihtelua aiheuttavat tekijät, kuten myös ennustamattomasti vaikuttavat tekijät. Näiden huomioonottamista varten on valmistauduttu huolellisella suunnittelulla ja järjestelmällisillä työskentelymetodeilla, jotka ovat resilienttejä myös odottamattomille muutoksille prosessin aikana. Tutkimuksen voidaan todeta olevan myös toteutettu tieteellisen tutkimuksen toteutusta yleisellä tasolla ohjaavien periaatteiden mukaisesti. Tutkimusta tehdessä on noudatettu Jyväskylän yliopiston ja hyvien tieteellisten käytänteiden noudattamista.

Viimeisenä luotettavuutta arvioitavana kohteena on tutkimuksen vakiintuneisuus. Tämä kytkeytyy löyhästi myös tutkimuksen luotettavuuden arviointiin, jota käsiteltiin edellä. Niin ikään ulkopuolisena henkilönä tämän pro gradu -tutkielman on arvioinut työn ohjaaja virallisen pro gradun -arviointikriteeristön mukaisesti.

TAULUKKO 24 Luotettavuuden kriteeristö laadullisessa tutkimuksessa (Tuomi & Sarajärvi, 2018, s. 162).

	Niranen (1990)	Tynjala (1991)	Eskola & Suoranta (1996)	Parkkila ym. (2000)
Credibility				
Uskottavuus	-	luotettavuus, jonka osatekiöitä ovat "totuusarvo", sovel-lettavuus, pysyvyys ja neutraalisuus	vastaavatko tutkijan tekemä käsitteellis-täminen a tulkinta tutkittavien käsityksiä	tutkimukseen osallistuneiden riittävä kuvaus ja arvio kerätyn aineiston totuu-denmukaisuu-desta
Vastaavuus	vastaavatko tutkijan tuottamat rekonstruk-tiot tutkittavien todelli-suudesta alkuperäisiä konstruktiota	vastaavatko tutkijan tuottamat rekonstruktiot tutkittavien todelli-suudesta alkuperäi-siä konstruktioita	-	-
Transferability				
Siirrettävyys	tulosten siirrettävyys toiseen kontekstiin riippuen siitä, miten samankaltaisia tutkittu ympäristö ja sovellus-ympäristö ovat	tulosten siirrettävyys toiseen kontekstiin rippuu siitä, miten samankaltaisia tutkittu ympäristö ja sovellusympäristö ovat	tulosten siirrettävyys toiseen kontekstiin mahdollista tietyin ehdoin, vaikka yleistykset eivät ole mahdollisia (sosiaa-lisen todellisuuden monimuotoisuuden vuoksi)	tulosten siirrettävyys tutkimuskontekstin ulkopuoliseen vastaavaan kontekstiin
Dependability				
Luotettavuus	ulkopuolinen henkilö tarkastaa tutkimuspro-sessin toteutumisen	-	-	-
Tutkimustilanteen arviointi	-	tutkijan tulee ottaa huomioon paitsi erilaiset ulkoiset vaihtelua aiheut-tavat tekijät, myös tutkimuksesta ja ilmiöstä itsestään johtuvat tekijät	-	-
Varmuus	-	-	tutkijan pitaä ottaa mahdollisuuksien mukaan huomioon tutkimukseen ennus-tamattomasti vaikut-tavat tekijat	-
Riippuvuus	-	-	-	tutkimus on toteutettu tieteellisen tutkimuksen toteuttamista yleisesti ohjaavin periaattein
Confirmability				
Vakiintuneisuus	ulkopuolinen henkilö arvioi tutkimuksen tuotokset (aineiston, löydökset, tulkinnat, suositukset ym.)	-	-	-

6.1.2 Arvio käytetystä kirjallisuudesta

Tässä pro gradu -tutkielmassa käytettyä kirjallisuutta arvioidaan tässä luvussa. Kirjallisuus teoreettista viitekehystä varten on etsitty pääasiallisesti hyödyntäen Jyväskylän yliopiston kirjaston JYKDOK-tietokantaa. Lisäksi tukena on käytetty Google Scholaria ja IEEE:n tietokantaa. Hakukoneita on myös hyödynnetty, joista käytössä ovat olleet Google sekä Bing.

Teoreettista viitekehystä varten on nojattu suurimmaksi osaksi alan tieteellisiin artikkeleihin, joita on julkaistu esimerkiksi erilaisissa IEEE:n julkaisuissa tai Springerin julkaisemissa kokonaisuuksissa. Lisäksi mukaan on otettu yksittäisiä artikkeleita. Niiden lisäksi mukana on perinteisiä kirjoja, jotka on lainattu Jyväskylän yliopiston kirjastosta. Kirjojen osalta mukaan on valittu kyberturvallisuusalalla ja tekoälyn tutkimuksessa tunnettuja teoksia. Tutkimusoppaat, joita on hyödynnetty, ovat suomalaisella korkeakoulukentällä tunnettuja ja paljon käytettyjä teoksia.

Tieteelliset artikkelit ovat pääasiassa vertaisarvioituja. Tämä valinta tehtiin, jotta voitaisiin kasvattaa tutkimuksen uskottavuutta ja luotettavuutta. Lisäksi tukena on käytetty muutamia verkkosivuja. Niihin on tukeuduttu niin vähän kuin mahdollista. Verkkosivut on valikoitu tarkkaan, jotta ne olisivat mahdollisimman luotettavia ja samaan aikaan relevantteja tämän tutkimuksen kannalta. Esimerkkinä mukana on IBM:n, F-Securen sekä CrowdStriken verkkosivuja. Kyseiset toimijat ovat alalla tunnettuja ja kansainvälisesti tunnustettuja toimijoita.

Käytetystä kirjallisuudesta voidaan esittää arvio, että se tukee sisällön ja relevanttiuden osalta tätä pro gradu -tutkielmaa hyvin. On pyritty käyttämään myös mahdollisimman uutta kirjallisuutta valikoiden käytetyt lähteet 2020-luvulta. Joitain lähteitä ja alan kivijalkateoksia ja -julkaisuja on toki myös kyseisen vuosikymmenen edeltävältä ajalta. Kritiikkinä käytetyn kirjallisuuden osalta voidaan samaan aikaan esittää. Teoreettista viitekehystä varten lähteitä olisi voinut olla vielä lisää, mutta tutkijan ajankäytön ja resurssienhallinnan näkökulmasta määrä, johon päädyttiin, oli paras mahdollinen.

6.1.3 Tulosten tarkastelu

Aineistolähtöisen sisällönanalyysin perusteella yritettiin löytää vastauksia tämän pro gradu -tutkielman tutkimuskysymyksiin. Pääkysymyksenä oli: ”Miten selittäviä tekoälymalleja voitaisiin hyödyntää kyberuhkien havaitsemisessa?”. Tukikysymyksinä olivat: ”Voidaanko selitettävillä tekoälymalleilla saavuttaa korkeampaa luotettavuutta ja tietoturvaa kyberturvallisuuden alalla?” sekä ”Kuinka selitettävät mallit soveltuisivat organisaatioille, jotka vaativat korkeaa laatua ja tarkkuutta käytettäviltä tietojärjestelmiltä?”

Kuten Tuomi ja Sarajärvi (2018, s. 117) mainitsevat, tähtää sisällönanalyysi siihen, että tutkittavasta ilmiöstä saataisiin kuvaus tiivistetyssä sekä samaan aikaan yleisessä muodossa. Samalla he huomauttavat, että sisällönanalyysi ei ole kuitenkaan avain valmiiseen tutkimukseen, vaan sen avulla saadaan ainoas-

taan tutkimusta varten kerätty aineisto järjestettyä, jonka pohjalta voidaan muodostaa johtopäätökset (Tuomi & Sarajärvi, 2018, s. 117).

Tämän alaluvun tavoitteena on pohtia ja peilata sisällönanalyysin tuloksia aikaisempaa tietoa sekä tämän tutkielman teoreettista viitekehystä vasten pitäen tutkimuskysymykset mukana. Kuten tuloksissa esiteltyä, analyysiprosessin ja tarkemmin viimeisen vaiheen, eli abstrahoinnin, tuloksena saatiin kolme yhdistävää luokkaa. Luokat olivat:

- XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa,
- perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan,
- XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla.

Ensimmäinen yhdistävä luokka koskee selittävien tekoälymallien kysyntää ja niiden tuomaa vahvistusta kyberturvallisuusratkaisuissa. Jos tarkastellaan ensin analyysin polkua taaksepäin tästä yhdistävästä luokasta, niin päästään ensin yläluokkiin:

- XAI:n tuomat edut ja mahdollisuudet kyberturvallisuuden kentälle,
- XAI:n paikka ja rooli kyberturvallisuuden ratkaisuissa.

Näiden luokkien taustalla on alaluokkia, joihin lukeutuu esimerkiksi luottamuksen vahvistaminen, tehokkuus, päätöksenteon tukeminen, XAI-mallien tarve, tekoälyn rooli sekä tekoälyn ja ihmisen välinen vuorovaikutus.

Kun tarkastellaan tutkielman teoreettista viitekehystä sekä analyysin tuloksia, niin voidaan melko selkeästi esittää väite siitä, että selitettävien tekoälymallien kysyntä kyberturvallisuuden alalla on ilmeinen. Samalla tuloksien ja aiemman tiedon valossa on nähtävissä, että nykyisin enemmän käytetyt koneoppimismallit, jotka ovat pääasiassa mustia laatikoita, eivät ole riittäviä luotettavuuden kannalta. Analyysiprosessia tehdessä kerätystä aineistosta nousi esiin useaan otteeseen ajatus siitä, että nykyisten tekoälymallien luotettavuuden, läpinäkymättömyyden ja ymmärrettävyyden puute aiheuttaa kestämättömän tilanteen kyberturvallisuuden ratkaisuissa.

Voidaan vetää tulkinta, että tekoälyä tulisi käyttää ihmisen assistenttina ja työkaluna, eikä niinkään kaikki tietävänä artefaktina, johon luotetaan sataprosenttiesti ja, jonka ennusteiden ja tulosten pohjalta tehtäisiin päätöksiä ilman ihmisen tuomaa ajattelua ja kontribuutiota. Samaan aikaan vaikuttaa siltä, että suorituskyky on monessa kohtaa kuitenkin hyvin tärkeä elementti, mutta siitä saatetaan sokaistua helposti.

Esimerkkinä; organisaation verkkoliikenteestä kyetään tekoälyn avulla havaitsemaan anomalia, jonka perusteella sisään tunkeutuneen uhkatoimijan jäljille päästään ja täten onnistutaan lieventämään vahinkoja sekä parhaassa tapauksessa ennaltaehkäisemään uusia tuhoja. Herkästi tekoälyä käyttävän organisaation kyberturvallisuuden asiantuntijat voivat esittää kysymyksen: ”Mihin selityksiä tarvitaan, jos käytetty tekoälymalli toimii tehokkaasti ja tarkasti?”

On selvää, että selitettävyyden lisääminen tekoälymalleihin tuo lisää kustannuksia. Samalla voidaan tulosten pohjalta todeta, että selitysmenetelmät ovat vielä vailla laaja-alaista tutkimusta ja kehitystä, jotka osiltaan nostavat kustannuksia merkittävästi.

Tullaankin ehkä tilanteeseen ja tienristeykseen, jossa vaakakupissa painavat myös moraaliset sekä eettiset kysymykset tekoälyn käytön suhteen, kun pohditaan selitettävyyden lisäämistä tekoälyyn. Voiko jokin artefakti todella olla luotettava, jos emme tiedä miten se toimii tai miksi se päätyi johonkin tulokseen, mihin se päätyi? Minkä arvon laskemme sille, että käyttämämme tietojärjestelmät ovat avoimia ja läpinäkyviä, jos käytämme niitä esimerkiksi sellaisten päätösten tukena, jossa vaakalaudalla saattaa olla ihmisen terveys tai jopa henki?

Seuraavassa esitettynä suoria lainauksia kerätystä tutkimusaineistosta, jotka ottavat kantaa XAI-mallien luottamukseen:

> XAI is revolutionizing different ML techniques by producing more explainable models while maintaining predictive accuracy. Transparency in system design has enabled the end-users to understand and trust the decisions of the AI systems. (Islam ym., 2022, s. 3.)

> In all application domains, establishing the trust in and fairness of machine learning systems matters most in low-risk environments. In contrast, robustness and reliability are critical to high-risk environments where machines take over decisions with far-reaching consequences. (Kabir ym., 2022, s. 252.)

> This ultimately leads to ambiguous situations where the system becomes incompetent to explain the internal procedures and the reasoning behind taking major and risky decisions of life [1]. Relying on a system that cannot explicitly explain the purpose of making a decision leaves question on the effectiveness of the system. (M. Ahmed & Zubair, 2022, s. 267.)

Myös tutkimusta varten tehdyssä asiantuntijahaastattelun vastauksissa esiin nousivat kysymykset luottamuksesta ja sen puutteesta tekoälymalleja kohtaan. Haastateltava nosti esiin huolen siitä, että mitä enemmän tekoälyä implementoidaan kriittisiin paikkoihin, sitä enemmän kysyntä selitettävyyttä ja mallien läpinäkyvyyttä kohtaan kasvaisi. Samalla haastateltava totesi, että jos käytetty tekoälymalli toimii, kiinnostus selityksiä kohtaan ei välttämättä ole kovin korkea. Hän myös pohti selitettävyyttä siitä näkökulmasta, että tietokoneiden ollessa ihmisten rakennelmia, niiden ei pitäisi olla milloinkaan selittämättömissä.

Hyvä ja tärkeä huomio oli myös haastateltavan näkemys siitä, kuinka koneoppimismallit tuottavat useasti vinoutunutta (engl. biased) sisältöä ja päätöksiä. Tästä näkökulmasta myös selitykset olisivat tärkeitä, koska muuten tapahtuu ajautuminen puolueelliseen päätöksentekoon.

Haastateltava esitti myös aiheellista kritiikkiä selitystekniikoita kohtaan. Jos ollaan tilanteessa, että ihmiset nojaavat tekoälyn tuottamien päätösten lisäksi selitysmallien antamiin selityksiin, voi se olla vaarallinen yhtälö. Perusteena tälle väitteelle haastateltava esitti pohdinnan siitä, kuinka selitystekniikatkin ovat luultavasti koneoppimismalleja. Mikäli selityksiä antavat mallit onnistu-

taan myrkyttämään pahalla datalla tai opettamaan vastaamaan väärin, seuraukset voivat olla vakavia.

Voidaankin todeta, että on tärkeää ottaa huomioon ja nostaa keskusteluun myös selitettävän tekoälyn haavoittuvuudet ja niiden manipuloiminen vihamielisessä tai rikollisessa tarkoituksessa. Selitysmallien, kuten LIME:n, haavoittuvuuksia on nostettu tutkimuksissa esiin, mutta haavoittuvuuksista puhutaan vielä vähän.

Toinen abstrahoinnin tuloksena syntynyt yhdistävä luokka kuvasi perinteisten koneoppimismallien lähes tavoittamattomiin kasvanutta takamatkaa verrattuna kilpajuoksussa uusia kyberhyökkäystyyppejä vastaan. Analyysipolkua takaisin päin mentäessä vastaan tulee avoimuuden ja tulkittavuuden puute, jotka liittyvät perinteisiin koneoppimismalleihin. Myös jatkuva toimintaympäristön ja kyberturvallisuuskentän muutostila sekä näihin vastaaminen olivat yksi yläluokka.

Luvussa 4.2 käsiteltiin uhkatoimijoiden hyödyntämää tekoälyä ja sitä, kuinka vihamieliset hyökkäävät osapuolet ovat implementoineet ja hyödyntäneet tekoälyä jo puolustavaan osapuoleen pitkän aikaa. Haasteellista on, että hyökkääjiä eivät koske mitkään säännöt. Yhtäältä esimerkiksi EU-alueella toimiessa Euroopan unionin sääntely liittyen tekoälyn käyttöön tulee lisääntymään lähivuosina.

Tästä konkreettisena esimerkkinä Euroopan unionin parlamentissa valmisteilla oleva EU AI Act, joka tulee olemaan maailman ensimmäinen kattava tekoälylaki. Parlamentin tärkein tavoite AI Act:n osalta on varmistua siitä, että EU:ssa käytettävät tekoälyjärjestelmät olisivat turvallisia, läpinäkyviä, jäljitettäviä, syrjimättömiä sekä ympäristöystävällisiä. Tavoitteiden mukaan tekoälyjärjestelmiä pitäisi valvoa automaation sijaan ihminen. (*EU AI Act*, 2023.)

Jos pohditaan tämän tutkimuksen tuloksia AI Act -lakia vasten, voidaan ehdottomasti todeta, että kyseiselle laille on suuri tarve. EU:lla on iso rooli esimerkin näyttäjänä koko maailmalle ja tekoälyn kehittäjille vastuullisena asetusten ja lakien säätäjänä. Ei voi kuitenkaan olla nostamatta esiin huolta, joka liittyy tuloksissa esiin nousseeseen kuiluun hyökkääjien ja puolustajien välillä kybermaailmassa, joka liittyy koneoppimisen käyttöön. Olisi tärkeää, että regulaatio ei hidastaisi liiaksi tekoälyn kehitystä, ja että sitä toteuttavilla yhtiöillä olisi riittävä vapaus tuoda uusia teknologioita organisaatioiden hyödynnettäväksi.

Kyberturvallisuusratkaisujen näkökulmasta olisi ehdottoman tärkeää, että kuilu uhkatoimijoiden ja puolustajien välillä ei ainakaan pääsisi enää kasvamaan. Tuloksista ilmenee kuitenkin, että nimenomaan kyberturvallisuusratkaisuissa käytettyjen tekoälymallien vaikea ymmärrettävyys ja läpinäkyvyyden puute ovat osatekijöitä puolustuksen takamatkan suhteen. Avoimuuden ja selitettävyyden lisääminen ei siis automaattisesti tarkoita vain kustannusten nousua ja kehityksen hidastumista, vaan samaan aikaan niillä on nähtävissä päinvastaiset vaikutukset.

Oikeudellisesta näkökulmasta tutkimuksen tuloksista nousi esiin kiinnostava pohdinta haasteltavalta liittyen tekoälyratkaisuihin kyberturvallisuuden ratkaisuissa. Haastateltava pohti tilannetta, jossa oikeuden tai jonkin kansainvälisen yhteisön edessä käsiteltäisiin tilannetta, jossa pitäisi pystyä esittämään todisteita kyberhyökkäyksen alkuperästä. Haasteltava esitti epäilyksen, että

miten voidaan uskottavasti perustella ratkaisua hyökkääjän alkuperästä, jonka koneoppimismalli on tehnyt, mutta samaan aikaan ei voida selittää miksi ja miten se on ratkaisuun päätynyt. Myös juridisesta näkökulmasta tarkasteltuna voitaisiin pohtia, tarjoaisiko selitettävyys hyötyjä.

Uhkatoimijoiden käyttämät hyökkäysvektorit, muuttuvat uhat sekä kybertoimintaympäristön muutos tuovat myös omat ongelmansa puolustautumisen näkökulmasta kybertoimintaympäristössä. Hyökkääjä harvoinkaan käyttää täysin samoja uhkatunnisteita tai tismalleen samoja hyökkäysvektoreita kovin useaan kertaan. Täten jo tapahtuneista ja paljastuneista kyberhyökkäyksistä saatu data ja sen pohjalta rakennettu tieto on useassa tapauksessa vanhaa, eikä siitä välttämättä ole paljoakaan hyödynnettävissä uusien hyökkäyksien estämisessä.

Näin mustavalkoinen ja yksinkertainen tilanne ei asian suhteen kuitenkaan ole. Jo tapahtuneiden hyökkäysten tutkinta ja niistä saadun tiedon pohjalta rakennettu tieto on tärkeää esimerkiksi kyberuhkatiedustelun näkökulmasta ja siihen nojaavien päätösten tekemisen suhteen. Silmiä ei voi kuitenkaan ummistaa siltä tosiseikalta, että hyökkäykset tulevat monimutkaistumaan tekoälyn myötä tulevaisuudessa. Kyberpuolustuksessa käytetyiltä tekoälyratkaisuilta odotetaan näin ollen nykyistä huomattavasti suurempia kykyjä ennakoida ja estää hyökkäyksiä. Selitettävyydellä on tässä yhtälössä iso rooli, jotta luotettavuus puolustusratkaisuihin olisi mahdollisimman korkealla tasolla.

Sisällönanalyysin kolmantena yhdistävänä luokkana muodostui selitettävien tekoälymallien tarpeen kasvu kyberturvallisuuden alalla. Tämä menee hieman päällekkäin ensimmäisen yhdistävän luokan kanssa. Voidaan tehdä johtopäätös ja esittää väite, että tekoälyn tarve kyberpuolustuksen ratkaisuissa on ilmeinen. Tekoälyn käyttö ei ole kuitenkaan lainkaan uusi keksintö kyberturvallisuuden ratkaisuissa. Vaaditaan kuitenkin tiedeyhteisöltä sekä yrityksiltä ja julkisen sektorin toimijoilta saumatonta yhteistyötä ja kaikkien osapuolten vahvaa kontribuutiota tutkimuksen tekemiseen selitysmenetelmien osalta. Tämän tutkimuksen tuloksista selviää myös, että tutkimushaasteita on myös esiintynyt.

Esimerkkinä on noussut esiin ristiriitoja siitä, mitä selitettävyys on tieteellisen ymmärryksen näkökulmasta. Tuloksien mukaan myös useiden tieteenalojen yhteen saattaminen tutkimuksen osalta on tärkeä seikka. Myös haasteltava toi esille hyvän lähestymistavan tekoälyn ja selitettävän tekoälyn tutkimiseen. Hän totesi, että kun koneoppimismalleja käytetään ratkaisemaan esimerkiksi yhteiskunnallisia ja ihmisten elämien kannalta merkityksellisiä asioita, tekoäly ei voi olla vain insinöörien kehittämä ratkaisu. Olisi todella tärkeää yhdistää mukaan sosiaalitieteitä ja muita tieteenaloja.

Tekoälyllä tulee olemaan monella tasolla todella suuri rooli jo lähitulevaisuudessa, joten poikkitieteellisen tutkimuksen voidaan todeta olevan hyvin tärkeä lähestymiskulma. Selitysmenetelmien näkökulmasta se on sitä ehdottomasti myös, koska loppujen lopuksi ihminen on se joka tekoälyn tekemiä päätöksiä ja ennusteita tulkitsee – ainakin toivottavasti.

Ei voi myöskään olla niin, että kriittisen infrastruktuurin kannalta merkittävissä organisaatioissa, kuten energiayhtiöissä tai hyvinvointialueilla, kyberturvallisuuden ratkaisuissa nojattaisiin vain koneoppimismalleihin, jotka ovat mustia laatikoita. Kybertoimintaympäristössä tehtävät ratkaisut tulevat muut-

tumaan entistä merkittävimmiksi useilla yhteiskunnan toiminnan kannalta kriittisillä toimialoilla, joten avoimuus ja ymmärrettävyys pitäisi pystyä implementoimana osaksi tekoälyratkaisuja jo suunnitteluvaiheessa.

Kiteytettynä erityisesti yhteiskunnan elintärkeiden toimintojen ja kriittisen infrastruktuurin organisaatioiden kyberturvaa rakennettaessa selitykset voisivat tämän tutkimuksen perusteella vankistaa ja lisätä käytettyjen tekoälymallien luotettavuutta. Läpinäkyvyyden parantuessa tekoälymalleihin nojautuva kyberturvallisuuteen liittyvä päätöksenteko voidaan perustella paremmin. Tarvittaessa voidaan myös jälkikäteen jäljittää tekoälyn tekemiä ratkaisuja tarpeen tullen. Täten voidaan varmistua paremmin siitä, että yhteiskunnat toimivat myös tilanteissa, joissa kyberhyökkäyksillä yritetään horjuttaa niitä. Lait ja asetukset koskettavat kriittisen infrastruktuurin organisaatioita hieman voimakkaammin, kuin esimerkiksi niiden ulkopuolelle lukeutuvia pieniä yrityksiä ja organisaatioita. Näin ollen selitettävyyden käyttöönotossa pitää huomioida käytettyjen selitettävien tekoälymallien kyberturva ja itse selitettävyyttä kohtaavat uhat ja manipuloinnin mahdollisuus. Tämä osiltaan todennäköisesti nostaa kriittisien infrastruktuurin organisaatioiden kynnystä implementoida selitettäviä malleja ilman kattavaa tutkimusta ja testaamista.

Vaikka on olemassa selitystekniikoita, joiden avulla voidaan muodostaa selityksiä jälkikäteen (engl. post-hoc), niin olisi myös tärkeää, että selitykset olisivat saatavilla tarvittaessa jo mallin suorituksen aikana. Asiat tapahtuvat kybermaailmassa ja tietoverkoissa aina vain nopeammin ja nopeammin ja havaintojen mukana myöskään hyökkääjät eivät paina jarrua, vaan päinvastoin. Myös selitysten muodostamisen nopeudella on tässä yhtälössä merkittävä rooli, kun päätöstentekijöiden pitää pystyä reagoimaan tulevaisuudessa entistä nopeammin. Samaan aikaan kyberhyökkäykset tulevat erittäin suurella todennäköisyydellä muovautumaan vakavimmiksi ja monimutkaisemmiksi.

Tutkimusta suunniteltaessa asetettiin rajaus, jonka mukaan pyrittäisiin rajaamaan tutkimusta selitettävien tekoälymallien hyödyntämiseen hyökkäysketjumallien ensimmäisiin vaiheissa. Samalla todettiin, että tämä voi kuitenkin olla haastavaa tarpeeksi spesifin lähdeaineiston puutteen vuoksi. Tämä olettamus osoittautui todeksi ja rajaus tästä näkökulmasta oli hyvin haastavaa. Tutkimuksen teoreettisen viitekehyksen sekä tulosten valossa on kuitenkin havaittavissa viitteitä siitä, että esimerkiksi tietojenkalastelun havaitsemisessa ja sittemmin estämisessä on havaittu käytettävän selitettäviä malleja. Tämä voitaisiin tulkita hiljaiseksi signaaliksi siitä, että hyökkäysketjumallien ensimmäisten vaiheiden näkökulmasta selitettävyyttä kannattaisi tutkia lisää ja XAI-sovelluksilla on potentiaalia etenkin tiedusteluvaiheen näkökulmasta. Tutkimuksessa selvisi samalla, että hyökkäävät osapuolet käyttävät tekoälyä paljon juuri ensimmäisissä vaiheissa. Tästä näkökulmasta katsottuna voisi olla aiheellista pohtia jatkotutkimusta tämän tutkimuksen aihetta vasten.

Lopuksi voidaan todeta myös, että selitettävyyden implementoinnissa kyberturvallisuuden tekoälyratkaisuihin, huomioon pitää ottaa sekä reaktiivisuus että proaktiivisuus. Yhtäältä pitää pystyä reagoimaan odottamattomiin tilanteisiin sekä kyetä tekemään mahdollisimman tehokkaasti ennakoivaa toimintaa, jotta kyberhyökkäykset eivät pääsisi toteutumaan.

6.1.4 Tutkimuksen haasteet

Tutkimusprosessin aikana ilmaantui muutamia haasteita, jotka avataan tässä. Tutkimusmenetelmän valinnan osalta osoittautui haasteelliseksi valita menetelmä, joka soveltuisi tämän kaltaisen uuden ja vielä jokseenkin vähän tutkitun aiheen tutkimiseksi. Haastavaa oli tehdä päätös, tuleeko tutkimus olemaan laadullinen vai määrällinen. Määrällisen tutkimuksen etuna olisi ollut esimerkiksi paremmat mahdollisuudet kehittää jokin oma selitysmalli ja raportoida sen teknisestä toiminnasta.

Laadullinen tutkimus ja edelleen aineistolähtöinen sisällönanalyysi valikoitui kuitenkin pitkän harkinnan jälkeen käytettäväksi tutkimusmenetelmäksi. Valinta tehtiin sillä perusteella, että haluttiin tutkia selitettävää tekoälyä kyberturvallisuuden kentällä uutena ja jopa hieman tuntemattomana ilmiönä ja löytää sitä kautta tutkijaa askarruttaneisiin kysymyksiin vastauksia.

Tutkimus on aika kapea eikä välttämättä onnistu antamaan kattavaa kuvaa tutkittavasta aiheesta ollen tarpeeksi luotettava ja toistettava. Olemassa olevan kirjallisen aineiston puitteissa voidaan kuitenkin todeta, että tässä pro gradu -tutkielmassa onnistutaan tuomaan uusia näkökulmia kyberturvallisuuden kentälle ja samalla herättämään aiheen ympärillä keskustelua.

Haasteena oli myös sopivan aineiston kerääminen, joka olisi tarpeeksi luotettavaa ja samalla relevanttia. Prosessin alkuvaiheessa sopivaa aineistoa tuntui olevan hyvin vaikea löytää, mutta lopulta onnistuttiin saamaan kohtuullisen kattava aineisto kasaan.

Analyysiin liittyi omat haasteensa, jotka kytkeytyivät enimmäkseen oleellisen tiedon löytämiseen ja rajaamiseen. Selitettävään tekoälyyn liittyvät ilmaisut menevät hyvin paljon päällekkäin muun tekoälyyn liittyvän tekstin ja keskustelun kanssa. Tämän takia analyysin aikana jouduttiin kamppailemaan jonkin verran, että onnistuttiin tuomaan esiin tämän tutkimuksen kannalta oleellinen tieto.

6.1.5 Tulevaisuuden tutkimuksen tarpeet

Kuten jo aiemmin tässä pro gradu -tutkielmassa on tullut esiin, tutkimuksen tarve selitettävien tekoälymallien osilta kyberturvallisuuden kentällä on kova. Tieteellisiä artikkeleita aiheesta on viimeisen kahden vuoden aikana ilmestymään enenevissä määrin, mutta silti aiheesta puhutaan vielä tutkijan havaintojen mukaan hyvin vähän. Tutkijan omat havainnot ja kokemus ovat myös samansuuntaisia. Monelle alan kyberturvallisuusalan asiantuntijallekin selittävä tekoäly on vielä hämärän peitossa oleva tai jopa tuntematon tekoälyn osa-alue.

Pelkästään EU:n uusi AI Act pakottaa tekoälyn kehittäjiä luomaan malleistaan avoimempia, mutta ei ole kestävää, jos se tulee olemaan ainoa ajuri tutkimuksen suhteen. Voidaan todeta, että selitettävyys myös kasvattaa tekoälymallien tarkkuutta ja tehokkuutta, joten myös nämä seikat todennäköisesti motivoivat myös uuden tutkimuksen tekemiseen. Myös kansalaisyhteiskunnilla ja erilaisilla yhteenliittymillä, kuten EU:lla, on iso rooli tutkimuksen eteenpäinviemisellä ja sen tarpeen korostamisella.

Tekoäly on ylikuumentunut aihe tällä hetkellä, ja on todellinen riski, että sen tuomien taloudellisten hyötyjen varjoon jäävät siihen sen muodostamat riskit. Etenkin kybertoimintaympäristössä tekoälyn implementoimisen pitää olla vastuullista ja eettisesti kestävää samoin kuin esimerkiksi terveydenhuollon piirissä ja kaikissa muissakin organisaatioissa, jotka ovat merkittäviä kriittisen infrastruktuurin kannalta. Selitettävyyden tutkimiseen pitkäjänteisesti panostaminen tulee todennäköisesti maksamaan itsensä monin verroin takaisin tulevaisuudessa.

Kuten tässäkin tutkimuksessa on tullut esiin, tulevaisuuden tutkimuksessa tärkeää on useiden eri tieteenalojen yhteen saattaminen. Tekoälyn vaikutukset tulevat olemaan yhteiskuntiimme ja koko ihmiskuntaan niin laaja-alaiset, että mikään yksi tieteenala ei kykene ratkaisemaan tutkimuskentän haasteita. Ja kuten mainittua, niin ihminen on hyvin keskeinen elementti tekoälyn käyttäjänä ja hyödyntäjänä, että esimerkiksi sosiaali- ja kognitiotieteillä tulee olemaan merkittävä rooli myös selitettävyyden tutkimisessa.

6.2 Johtopäätökset

Tässä pro gradu -tutkielmassa on tutkittu selitettäviä tekoälymalleja kyberuhkien havainnoinnissa. Tutkijan kiinnostus aihetta kohtaan alkoi kasvaa jo noin vuosi ennen pro gradu -tutkielman tekemisen aloittamista. Tutkimuksen tavoitteena oli etsiä ilmiöitä ja merkityksiä selitettävän tekoälyn ympäriltä kyberuhkien havainnoinnin kontekstissa. Samalla pyrittiin rakentamaan ymmärrystä siitä, miten tekoälyä hyödynnetään nykyään kybertoimintaympäristössä ja kuinka selitettävyys siellä asemoituu. Tavoitteena oli myös selvittää, voitaisiinko selitettävyydellä tehdä tekoälyn käytöstä luotettavampaa ja tarkempaa kyberturvallisuuden sovelluksissa ja ratkaisuissa. Lisäksi yksi alatutkimuskysymys oli aseteltu siitä näkökulmasta, kuinka selitettävät tekoälymallit soveltuisivat kriittisen infrastruktuurin organisaatioille niiden kyberturvaa rakennettaessa.

Tutkimuksen teoreettinen viitekehys ja tieteellinen pohja luotiin kirjoittamalla kirjallisuuskatsaus. Siinä käsiteltiin kyberturvallisuuden toimintaympäristöä sekä tekoälyn hyödyntämistä kyberturvallisuusympäristössä sekä hyökkäävän että puolustavan osapuolen näkökulmasta. Lisäksi käsiteltiin selitettäviä tekoälymalleja yleisesti sekä niiden hyödyntämistä kyberturvallisuuden ratkaisuissa.

Tutkimus toteutettiin laadullisena tutkimuksena ja tutkimusmenetelmäksi valikoitui aineistolähtöinen sisällönanalyysi. Tutkimuksen aineisto koostui tieteellisistä artikkeleista, jonka lisäksi aineistoa tukemana toteutettiin yksi asiantuntijahaastattelu. Aineistoon kerätyt artikkelit olivat vertaisarvioituja. Käytettiin myös ainoastaan kansainvälisiä julkaisuja. Haastattelu toteutettiin struktuuroimattomana syvähaastatteluna. Tutkimus toteutettiin loppuvuoden 2023 ja kevään 2024 aikana. Sisällönanalyysin tuloksena muodostui kolme yhdistävää luokkaa:

- XAI-mallien kysyntä ja niiden tuoma vahvistus kyberturvallisuuden tekoälyratkaisuissa,
- perinteisten ML-mallien lähes tavoittamattomiin kasvanut takamatka kilpajuoksussa uusia hyökkäystyyppejä vastaan,
- XAI-mallien tutkimuksen tarpeen kasvu kyberturvallisuuden alalla.

Aineistolähtöisen sisällönanalyysin rajoitteeksi ja haasteeksi muodostui relevantin aineiston niukkuus, koska tutkittava aihe on vielä uudehko. Aineisto rajattiin siten, että etsittiin vain sellaista aineistoa, joka koskee selitettäviä tekoälymalleja nimenomaisesti kyberturvallisuuden kentällä. Aineistoon valikoitui kuitenkin myös artikkeleita, jotka kytkeytyvät tutkimuskysymykseen hieman löyhemmin. Rajaus tehtiin myös julkaisuvuoden suhteen ja pois jätettiin kaikki vanhemmat kuin 2021 julkaistut artikkelit.

Tutkimuksen tuloksista on vedettävissä yhtäläisyyksiä verrattuna aiempaan tutkimukseen. Myös tutkimuksessa muodostettu keskustelu ja pohdinta tukevat aiempaa tutkimusta. On selkeästi havaittavissa, että tekoälymallien avoimuudelle, tulkittavuudelle ja selitettävyydelle on tarve kyberturvallisuuden kentällä. Selitettävyys lisää tekoälyn luotettavuutta ja auttaa havainnoimaan kyberuhkia. Selitettävyys myös madaltaa tekoälymallien implementoinnin kynnystä kyberpuolustuksen sovelluksissa, koska sen avulla tekoälyn tekemiä päätöksiä voidaan ymmärtää paremmin ja täten vastata uhkiin tehokkaammin.

Selitettävillä tekoälymalleilla voitaisiin myös parantaa kriittisen infrastruktuurin organisaatioiden kyberturvaa. Tämä vaatii kuitenkin kattavaa tutkimusta ja testaamista.

Tulevaisuuden tutkimukselle katsotaan myös olevan merkittävä tarve. Selitystekniikoita ei ole vielä tutkittu kyberturvallisuuden alalla kovin laaja-alaisesti, vaikka laadukkaita avauksia kansainvälisen tieteen kentällä on kuitenkin jo nähty. Esimerkiksi IEEE:ssä on jo muutamia julkaisuja aiheeseen liittyen. Myös esimerkiksi EU:n regulaatio ja valmisteilla oleva AI Act tulee asettamaan omat haasteensa ja paineensa tutkimukselle, mutta voidaan todeta, että AI Act:lle on tarve.

Tutkimuksessa tärkeänä esiin noussut havainto oli lisäksi ihmisen rooli selitettävien tekoälyratkaisujen osatekijänä. Ihminen on kuitenkin lopulta se, kuka myös selitettävät mallit kehittävät ja niitä myös käyttää kyberturvallisuuden kentällä. On siis tärkeää ottaa tulevaisuuden tutkimuksessa huomioon, että tutkimuksen pitäisi olla poikkitieteellistä huomioiden esimerkiksi sosiaali- ja kognitiotieteet. Lopuksi todettakoon, että tekoäly on hyvä renki, mutta huono isäntä.

LÄHTEET

Ahmed, I. (2022). From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, *18*(8), 5031. https://doi.org/10.1109/TII.2022.3146552

Ahmed, M., Islam, S. R., Anwar, A., Moustafa, N., & Pathan, A.-S. K. (Toim.). (2022). *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence* (Vsk. 1025). Springer International Publishing. https://doi.org/10.1007/978-3-030-96630-0

Ahmed, M., & Zubair, S. (2022). Explainable Artificial Intelligence in Sustainable Smart Healthcare. Teoksessa M. Ahmed, S. R. Islam, A. Anwar, N. Moustafa, & A.-S. K. Pathan (Toim.), *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence* (ss. 265–280). Springer International Publishing. https://doi.org/10.1007/978-3-030-96630-0_12

Ajankohtaisia kysymyksiä ja vastauksia kriittisestä infrastruktuurista ja varautumisesta – Huoltovarmuuskeskus. (ei pvm.). Noudettu 13. tammikuuta 2024, osoitteesta https://www.huoltovarmuuskeskus.fi/a/ajankohtaisia-kysymyksia-ja-vastauksia-kriittisesta-infrastruktuurista-ja-varautumisesta/

Ali, A., Khan, M. A., Farid, K., Akbar, S. S., Ilyas, A., Ghazal, T. M., & Al Hamadi, H. (2023). The Effect of Artificial Intelligence on Cybersecurity. *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1–7. https://doi.org/10.1109/ICBATS57792.2023.10111151

André, J.-C. (2019). *Industry 4. 0: Paradoxes and Conflicts*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/jyvaskyla-ebooks/detail.action?docID=5825593

Andress, J., Ablon, L., Winterfeld, S., & Steve Winterfeld. (2014). *Cyber warfare: Techniques, tactics and tools for security practitioners*. Syngress.

Asiantuntijahaastattelu. (2024, helmikuuta 16). [Henkilökohtainen viestintä].

Ceron, R. (2019, joulukuuta 5). AI, machine learning and deep learning: What's the difference? *IBM Blog*. https://www.ibm.com/blog/ai-machine-learning-and-deep-learning-whats-the-difference/

Chai, Y., Zhou, Y., Li, W., & Jiang, Y. (2022). An Explainable Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 790–803. https://doi.org/10.1109/TDSC.2021.3119323

Chakraborty, A. (2022). Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation. *arXiv.Org*. https://doi.org/10.48550/arxiv.2209.13454

Charmet, F., Tanuwidjaja, H. C., Ayoubi, S., Gimenez, P.-F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., & Zhang, Z. (2022). Explainable artificial

intelligence for cybersecurity: A literature survey. *Annals of Telecommunications*, *77*(11), 789–812. https://doi.org/10.1007/s12243-022-00926-7

Clark, D., Berson, T., Lin, H. S., & Council, N. (2014). *At the nexus of cybersecurity and public policy: Some basic concepts and issues* (s. 150). https://doi.org/10.17226/18749

Clark, D. D. (2010). *Characterizing cyberspace: Past, present and future* [Working Paper]. © Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/141692

Conti, M., Dargahi, T., & Dehghantanha, A. (2018). Cyber Threat Intelligence: Challenges and Opportunities. Teoksessa A. Dehghantanha, M. Conti, & T. Dargahi (Toim.), *Cyber Threat Intelligence* (ss. 1–6). Springer International Publishing. https://doi.org/10.1007/978-3-319-73951-9_1

Cyber Kill Chain®. (ei pvm.). Lockheed Martin. Noudettu 20. maaliskuuta 2024, osoitteesta https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

EU AI Act: First regulation on artificial intelligence. (2023, kesäkuuta 8). Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

Exploring the opportunities and limitations of current Threat Intelligence Platforms. (ei pvm.). [Report/Study]. ENISA. Noudettu 3. toukokuuta 2024, osoitteesta https://www.enisa.europa.eu/publications/exploring-the-opportunities-and-limitations-of-current-threat-intelligence-platforms

Frilander, K. (2018). Kybermaailma ja kansainvälinen oikeus. *Viestimies*, *2*, 21–24.

Garnham, A. (2018). *Artificial intelligence: An introduction*. Routledge.

Gordan, M., Ghaedi, K., Saleh, V., Gordan, M., Ghaedi, K., & Saleh, V. (2023). *Industry 4.0 – Perspectives and Applications*. https://doi.org/10.5772/intechopen.100660

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254

Gulmezoglu, B. (2022). XAI-Based Microarchitectural Side-Channel Analysis for Website Fingerprinting Attacks and Defenses. *IEEE Transactions on Dependable and Secure Computing*, *19*(6), 4039–4051. https://doi.org/10.1109/TDSC.2021.3117145

Hirsjärvi, S., Remes, P., Sajavaara, P., & Sinivuori, E. (2009). *Tutki ja kirjoita* (15. uud. p). Tammi.

Huang, Y.-T., Lin, C. Y., Guo, Y.-R., Lo, K.-C., Sun, Y. S., & Chen, M. C. (2022). Open Source Intelligence for Malicious Behavior Discovery and Interpretation. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 776–789. https://doi.org/10.1109/TDSC.2021.3119008

Hyppönen, M. (2022). *Internet*. WSOY.

Islam, M. U., Mozaharul Mottalib, Md., Hassan, M., Alam, Z. I., Zobaed, S. M., & Fazle Rabby, Md. (2022). The Past, Present, and Prospective Future of XAI: A Comprehensive Review. Teoksessa M. Ahmed, S. R. Islam, A.

Anwar, N. Moustafa, & A.-S. K. Pathan (Toim.), *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence* (ss. 1–29). Springer International Publishing. https://doi.org/10.1007/978-3-030-96630-0_1

Kabir, M. H., Hasan, K. F., Hasan, M. K., & Ansari, K. (2022). Explainable Artificial Intelligence for Smart City Application: A Secure and Trusted Platform. Teoksessa M. Ahmed, S. R. Islam, A. Anwar, N. Moustafa, & A.-S. K. Pathan (Toim.), *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence* (ss. 241–263). Springer International Publishing. https://doi.org/10.1007/978-3-030-96630-0_11

Kaplan, J. (2016). *Artificial Intelligence*. Oxford University Press.

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Klimburg, A., & Mirtl, P. (2012). *Cyberspace and governance – A primer* (Vsk. 65). Österreichisches Institut für Internationale Politik (oiip).

Kuppa, A., & Le-Khac, N.-A. (2021). Adversarial XAI Methods in Cybersecurity. *IEEE Transactions on Information Forensics and Security*, *16*, 4924–4938. https://doi.org/10.1109/TIFS.2021.3117075

Kyberturvallisuuden sanasto – Turvallisuuskomitea. (2018, lokakuuta 3). https://turvallisuuskomitea.fi/kyberturvallisuuden-sanasto/

Kyberturvallisuus ja kybertoimintaympäristö. (ei pvm.). Ulkoministeriö. Noudettu 3. joulukuuta 2023, osoitteesta https://um.fi/kyberturvallisuus-ja-kybertoimintaymparisto

Kyberturvallisuusstrategia. (ei pvm.). Valtiovarainministeriö. Noudettu 7. joulukuuta 2023, osoitteesta https://vm.fi/kyberturvallisuusstrategia

Laari, T., Flyktman, J., Härmä, K., Timonen, J., & Tuovinen, J. (2019). *#kyberpuolustus: Kyberkäsikirja Puolustusvoimien henkilöstölle* [D5 Oppikirja, ammatillinen käsi- tai opaskirja taikka sanakirja]. Maanpuolustuskorkeakoulu. https://www.doria.fi/handle/10024/173254

Lehto, M. (2022). APT Cyber-attack Modelling: Building a General Model. *International Conference on Cyber Warfare and Security*, 121–129.

Lehto, M., & Neittaanmäki, P. (Toim.). (2022). *Cyber security: Critical infrastructure protection*. Springer. https://doi.org/10.1007/978-3-030-91293-2

Lemay, A., Calvet, J., Menet, F., & Fernandez, J. M. (2018). Survey of publicly available reports on advanced persistent threat actors. *Computers & Security*, *72*, 26–59. https://doi.org/10.1016/j.cose.2017.08.005

Li, H., Wu, J., Xu, H., Li, G., & Guizani, M. (2022). Explainable Intelligence-Driven Defense Mechanism Against Advanced Persistent Threats: A Joint Edge Game and AI Approach. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 757–775. https://doi.org/10.1109/TDSC.2021.3130944

Menahem, E. (2009). Improving malware detection by applying multi-inducer ensemble. *Computational Statistics & Data Analysis*, *53*(4), 1483–1494. https://doi.org/10.1016/j.csda.2008.10.015

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Mikä on kyberhyökkäys? | F-Secure. (ei pvm.). Noudettu 17. huhtikuuta 2024, osoitteesta https://www.f-secure.com/fi/articles/what-is-a-cyber-attack

MITRE ATT&CK®. (ei pvm.). Noudettu 20. maaliskuuta 2024, osoitteesta https://attack.mitre.org/

NIST SP 800-115. (2020). *NIST*. https://www.nist.gov/privacy-framework/nist-sp-800-115

Ozkan-Okay, M., Akin, E., Aslan, Ö., Kosunalp, S., Iliev, T., Stoyanov, I., & Beloev, I. (2024). A Comprehensive Survey: Evaluating the Efficiency of Artificial Intelligence and Machine Learning Techniques on Cyber Security Solutions. *IEEE Access*, *12*, 12229–12256. https://doi.org/10.1109/ACCESS.2024.3355547

Russell, S. J., Chang, M.-W., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J. M., Mansinghka, V., Norvig, P., Pearl, J., & Wooldridge, M. (2022). *Artificial intelligence: A modern approach* (Fourth edition). Pearson.

Sailio, M., Latvala, O.-M., & Szanto, A. (2020). Cyber Threat Actors for the Factory of the Future. *Applied Sciences*, *10*(12), 4334-. https://doi.org/10.3390/app10124334

Saleh, H., & Sen, S. (2018). *Machine Learning Fundamentals*. Packt Publishing.

Samtani, S., Chen, H., Kantarcioglu, M., & Thuraisingham, B. (2022). Explainable Artificial Intelligence for Cyber Threat Intelligence (XAI-CTI). *IEEE Transactions on Dependable and Secure Computing*, *19*(4), 2149–2150. https://doi.org/10.1109/TDSC.2022.3168187

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. *SN Computer Science*, *2*(3), 173. https://doi.org/10.1007/s42979-021-00557-0

Schaller, R. R. (1997). Moore's law: Past, present and future. *IEEE Spectrum*, *34*(6), 52–59. https://doi.org/10.1109/6.591665

Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. *Engineering Applications of Artificial Intelligence*, *94*, 103770. https://doi.org/10.1016/j.engappai.2020.103770

Sharikov, P. (2018). Artificial intelligence, cyberattack, and nuclear weapons-A dangerous combination. *Bulletin of the Atomic Scientists*, *74*(6), 368–373. https://doi.org/10.1080/00963402.2018.1533185

Spyrou, E. D., & Kappatos, V. (2023). XAI using SHAP for Outdoor-to-Indoor 5G Mid-Band Network. *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*, 862–866. https://doi.org/10.1109/CSNT57126.2023.10134625

Stallings, W. (2019). *Effective cybersecurity: Understanding and using standards and best practices*. Addison-Wesley.

Suomen kyberturvallisuusstrategia 2019 – Turvallisuuskomitea. (ei pvm.). Noudettu 7. joulukuuta 2023, osoitteesta https://turvallisuuskomitea.fi/suomen-kyberturvallisuusstrategia-2019/

Tekoälyn mahdollistamat kyberhyökkäykset. (2022, joulukuuta 13). Traficom. https://www.traficom.fi/fi/julkaisut/tekoalyn-mahdollistamat-kyberhyokkaykset

Trump's Pick for NSA/CyberCom Chief Wants to Enlist AI For Cyber Offense. (2018, tammikuuta 10). Nextgov.Com. https://www.nextgov.com/artificial-intelligence/2018/01/how-likely-next-nsacybercom-chief-wants-enlist-ai/145085/

Tuomi, J., & Sarajärvi, A. (2018). *Laadullinen tutkimus ja sisällönanalyysi* (Uudistettu laitos). Kustannusosakeyhtiö Tammi.

Turtiainen, H., Costin, A., Polyakov, A., & Hämäläinen, T. (2023). Offensive Machine Learning Methods and the Cyber Kill Chain. Teoksessa T. Sipola, T. Kokkonen, & M. Karjalainen (Toim.), *Artificial Intelligence and Cybersecurity: Theory and Applications* (ss. 125–145). Springer International Publishing. https://doi.org/10.1007/978-3-031-15030-2_6

Valecha, R., Mandaokar, P., & Rao, H. R. (2022). Phishing Email Detection Using Persuasion Cues. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 747–756. https://doi.org/10.1109/TDSC.2021.3118931

Ventre, D. (2020). *Artificial Intelligence, Cybersecurity and Cyber Defense*.

Vähäkainu, P., Lehto, M., & Neittaanmäki, P. (2018). *Tekoäly ja kyberturvallisuus: Raportti*. Jyväskylän yliopisto.

What is a Threat Actor? | IBM. (2024, huhtikuuta 12). https://www.ibm.com/topics/threat-actor

What is AI in Cybersecurity? | AI Cybersecurity Explained. (ei pvm.). Noudettu 14. huhtikuuta 2024, osoitteesta https://www.sophos.com/en-us/cybersecurity-explained/ai-in-cybersecurity

What Is Machine Learning (ML)? (2020, kesäkuuta 26). *UCB-UMT*. https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/

What Is Machine Learning (ML)? | IBM. (2024, huhtikuuta 5). https://www.ibm.com/topics/machine-learning

What is the Cyber Kill Chain? Introduction Guide - CrowdStrike. (ei pvm.). Crowdstrike.Com. Noudettu 20. maaliskuuta 2024, osoitteesta https://www.crowdstrike.com/cybersecurity-101/cyber-kill-chain/

Wiafe, I., Koranteng, F. N., Obeng, E. N., Assyne, N., Wiafe, A., & Gulliver, S. R. (2020). Artificial Intelligence for Cybersecurity: A Systematic Mapping of Literature. *IEEE Access*, *8*, 146598–146612. https://doi.org/10.1109/ACCESS.2020.3013145

Zhang, Z., Hamadi, H. A., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research. *IEEE Access*, *10*, 93104–93139. https://doi.org/10.1109/ACCESS.2022.3204051

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances,

challenges, and opportunities. *The Artificial Intelligence Review*, *55*(2), 1029–1053. https://doi.org/10.1007/s10462-021-09976-0