Tekoäly kyberturvallisuudessa

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Kesäkuu 2026
Tommi Salo

TURUN YLIOPISTO
Tietotekniikan laitos

Tommi Salo: Tekoäly kyberturvallisuudessa

LuK-tutkielma, 24 s.
Tietojenkäsittelytiede
Kesäkuu 2026

Maailmassa jatkuvan digitalisaation vuoksi myös kyberturvallisuuteen liittyvien toimenpiteiden on tarve kehittyä. Kyberhyökkäysten määrän kasvaminen kasvattaa turvavalvomoiden taakkaa ja tähän ratkaisuna voisi olla tekoäly ja tämän monet koneoppimismenetelmät. Tässä tutkielmassa tarkastellaan tekoälyn soveltamista kyberturvallisuudessa sekä siihen liittyviä olennaisia menetelmiä sekä haasteita. Tutkielman kohteina ovat yleisimmät tekoälyn soveltamisen muodot uhkien tunnistuksessa sekä tämän automaattisen toiminnan käyttö. Tutkimme tarkemmin verkkohyökkäysten, haittaohjelmien ja tietojenkalastelun yritysten tunnistamista ja mitkä tekoälynmallit voivat tukea missäkin tehtävissä. Kirjallisuuskatsauksen lopussa pohditaan tekoälyn soveltamisen muotoja ja kuinka kaikkia tekoälyn soveltamisen muotoja ei vielä ole keksitty. Pohdinnassa käydään myös läpi generatiivisen tekoälyn tuomia huolia tulevaisuudessa.
Tutkimus toteutettiin kirjallisuuskatsauksena, jossa hyödynnettiin ajankohtaisia tieteellisiä julkaisuja. Työn tulosten perusteella voimme todeta tekoälyn tarjoavan kyberturvallisuudelle ympäristöönsä sopeutuvan toimijan, mutta tämän soveltaminen tuo mukanaan myös mahdollisia riskejä jotka rajoittavat tämän luotettavuutta. Tekoälyn soveltaminen edellyttää tekoälymallien luotettavuuden lisääntyvän varsinkin luomalla tekoälymalleista kestävämpiä manipuloivia hyökkäyksiä vastaan ja keksimällä ratkaisuja yksityisyyden takaamiseen. Työn jatkotutkimukseksi todettiin tarve tarkemmalle katsaukselle tietystä tekoälyn soveltamisen muodosta tai tutkielma keskittyen vain soveltamisen haasteisiin liittyvien ratkaisujen löytämiseen.

Asiasanat: kyberturvallisuus, tekoäly, verkkohyökkäykset, tietojenkalastelu, haittaohjelmat

Sisällys

Kuvat

Taulukot

1 Johdanto

Kyberhyökkäyksien määrä on kasvanut viimevuosien aikana maailmanlaajuisesti. Perinteiset tunkeutumisen havaitsemisjärjestelmät (engl. Intrusion Detection System, IDS) eivät kykene suojelemaan uusilta kyberuhkatoimijoilta (engl. Advanced Threat Actor, APT) tai kiristyshaittaohjelmilta [1]. Tämän lisäksi kyberturvaratkaisut kuten palomuurit, virustorjuntaohjelmat ja tunkeutumisen havaitsemisjärjestelmät eivät välttämättä riitä torjumaan monimutkaisia sekä yleistyviä kyberuhkia [2].

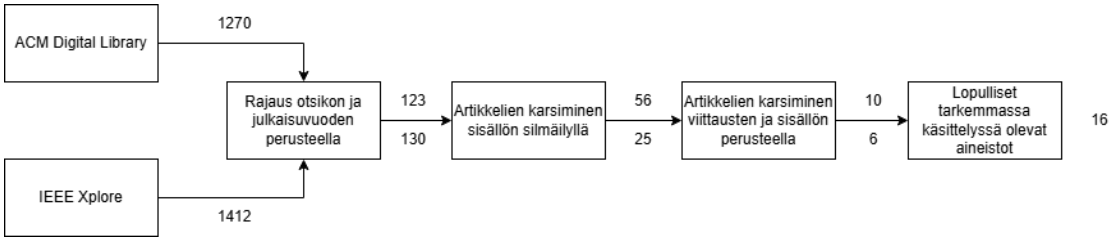
Tekoäly (engl. artificial intelligence) sekä koneoppiminen (engl. machine learning) kykenevät analysoimaan suuria tietokantoja sekunneissa kuten myös havaitsemaan kuvioita ja ilmeneviä uhkia. Tekoäly ja varsinkin syväoppiminen (engl. deep learning) on kehittynyt havaitsemaan kuvia sekä ihmisten ääniä. Nämä mallit ovat myös kehittyneet havaitsemaan haittaohjelmia sekä verkkomurtoja paremmin kuin ennen. [1]

Tämän tutkielman tavoite on selvittää, kuinka tekoälyä voidaan soveltaa kyberturvallisuudessa. Tutkimme keinoja parantaa kyberturvallisuutta tekoälyn avulla ja mitä haasteita tekoälyn käytössä voi esiintyä. Tutkielman tavoitteiden pohjalta olen jäsennellyt seuraavat tutkimuskysymykset:

TK1: Miten voimme hyödyntää tekoälyä kyberturvallisuudessa?

TK2: Mitä haasteita tekoälyn soveltamisessa on?

Tutkielma on toteutettu kirjallisuuskatsauksena. Aineistojen keräämiseen on käytetty kahta tietokantaa: IEEE Xplore ja ACM Digital Library. Keskeisimmät hakulauseet ovat olleet abstraktissa ("artificial intelligence"OR "AI"OR "machine learning") AND ("cybersecurity"OR "cyber security"OR "information security") AND ("improvement"OR "protection"OR "threat detection"OR "prevention") ja otsikoissa ("artificial intelligence"OR "AI"OR "machine learning"). Hakulauseen rajaaminen abstraktiin sekä otsikkoon karsi pois artikkeleita, jotka eivät sisältäneet aiheeseen liittyvää tietoa näiden tiivistyksissä ja otsikoissa. Tämän lisäksi lähteet rajattiin vuosiin 2022–2026, jotta tutkielma hyödyntää mahdollisimman tuoretta tietoa tekoälyn ollessa nopeasti kehittyvä aihepiiri. Aineistossa painotetaan monta lähdettä, joihin monet muut ovat tutkimuksissaan viitanneet. Tämä takaa lähteiden luotettavuuden. Lähdehakuprosessia kuvastaa Kuva 1.1.

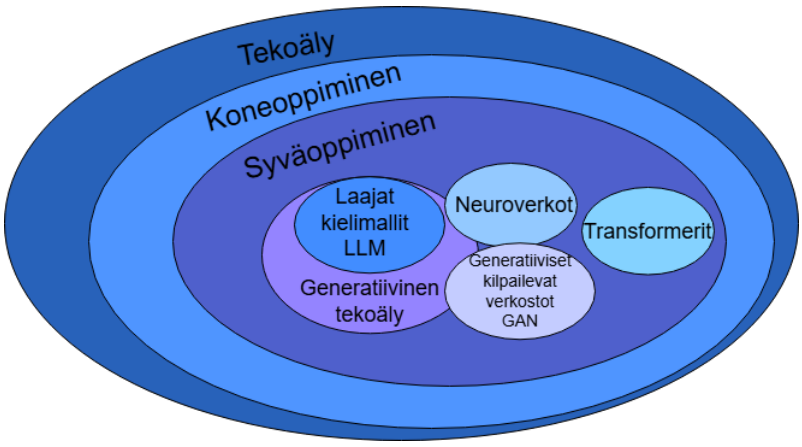


Kuva 1.1: Aineistojen haku

ACM Digital Library toi minulle 1270 hakutulosta kun hain vain abstraktissa olevilla hakulauseilla. Rajaamalla otsikon ja julkaisuvuoden perusteella sain 123 artikkelia, joista karsin pois artikkelien sisältöä silmäillen suuren osan jättäen vain 56 artikkelia. Lopuksi otin tarkempaan tarkasteluun viittausten ja artikkelien sisällön perusteella vain 10. IEEE Xplore toi abstraktin hakulauseella 1412 hakutulosta, joista rajaamalla otsikon ja julkaisuvuoden perusteella jäi vain 130. Näistä poistin artikkelit, joihin en päässyt käsiksi. Karsin myös silmämääräisesti turhat artikkelit, jotka eivät soveltuneet tutkimukseeni. Tämä jätti minulle 25 joista käsittelyyn otin viittausten ja sisällön perusteella lopuksi 6. Tällöin lopulliseen tarkasteluun valitsin vain 16 lähdettä, joista tein tutkimuksen.

2 Tekoälyn perusteet kyberturvallisuudessa

Tekoälyllä voidaan tarkoittaa useampaa eri tekoälyn mallia ja näiden erittely on tärkeää [3] sillä vuoden 2023 lopussa kansan suosioon tullut laajan kielimallin tekoäly (engl. large language model, LLM) ChatGPT on yksi tunnetuimpia tekoälyn muotoja [4], mutta tämä ei kuvasta tekoälyä kokonaisuudessaan. Tekoäly on teknologian muoto, jonka tarkoitus on luoda kone, jolla voidaan imitoida ihmisen aivojen toimintaa. Tekoälyn tavoite on pystyä toteuttamaan toimenpiteitä mihin vain ihmisen aivot pystyisivät kuten päätöksen tekoa, ongelmanratkaisua sekä kielen ymmärtämistä, että oppimista.[5][6]. Suurin osa tekoälymalleista pohjautuu koneoppimiseen ja tämän alaluokkiin [3] ja näitä kuvastaa Kuva 2.1.



Kuva 2.1: Kuvastaa tekoälystä olevia muotoja, joka pohjautuu [3] kuvaan

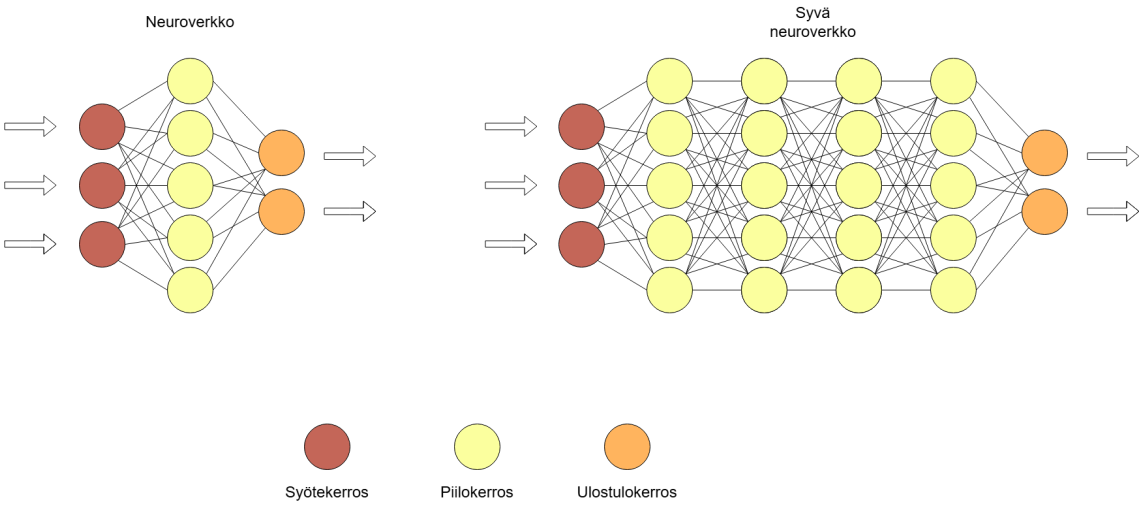
Keskeiset menetelmät kyberturvallisuudessa. Kyberturvallisuus on alana kokenut huomattavaa muutosta tekoälyn yleistymisen myötä. Tämä muutos on tapahtunut koneoppimisen, syväoppimisen, luonnollisen kielen käsittelyn (engl. natural language processing, NLP) poikkeamantunnistusalgoritmien (engl. anomaly detection algorithms) ja muiden tekoälymallien ansiosta. Nimettyjä tekoälymalleja käytetään näiden kyvystä sopeutumaan tilanteeseen sekä turvaamistehtävään. [2]

Koneoppiminen ja syväoppiminen. Koneoppimisalgoritmi on ”oppiva algoritmi” eli laskennallinen prosessi, joka hyödyntää syötettyä dataa saavuttamaan tahdotun tehtävän lopputuloksen. Näitä malleja ei ohjelmoida saavuttamaan tietty lopputulos vaan algoritmit ovat ”pehmeästi määritelty” eli voivat mukauttaa rakennettaan automaattisesti toistojen kautta ollakseen tehokkaampia. [7] Koneoppimisen tavoitteena on luoda koneita, joilla on kyky automaattisesti tehdä omat päätöksensä [8]. Näitä malleja koulutetaan koulutusvaiheessa materiaalin avulla, samalla kun malleille voidaan ilmaista haluttu lopputulos. Tämän jälkeen koneoppimismallit pystyvät itsenäisesti tunnistamaan haluttuja piirteitä uudesta datasta, jota tälle syötetään. Ideaalisen mallin on tarkoitus pystyä emuloimaan ihmisen kykyä prosessoida tietoa itsenäisesti ja sopeutua tälle annettuun tehtävään ja ympäristöön. [7] Näitä malleja voidaan kouluttaa valvotulla tai valvomattomalla oppimisella [8][7].

Valvotussa oppimisessa (engl. supervised learning) koulutusmateriaali sisältää piirteitä (väri, muoto, tekstuurit yms.) ja oikean luokan, joihin nämä luokitellaan. Tämä mahdollistaa tarkan luokittelun etenkin, kun luokiteltavassa objektissa esiintyy useita piirteitä. Tämä mahdollistaa koulutettavan mallin tunnistamaan objekteja, joilla on myös muuttuvia piirteitä omissa luokissa ja osaa erotella eri luokat toisistaan. Ei valvotussa oppimisessa eli valvomattomassa oppimisessa (engl. unsupervised learning) tavoitteena on antaa koneoppimismallille vapautta saavuttaa tavoite omin keinoin. Koulutuksessa ei anneta datan mukana ohjeita tai suuntaa, vaan tarkoituksena on saada malli itsenäisesti löytämään piirteet sekä tavoitteen datas-

ta. Koulutuksen aikana mallin vapauksia rajataan jokaisen yrityksen jälkeen, jotta malli pääsee lähemmäs tavoitetta. [8][7] Oppimista voidaan tukea vahvistusoppimisella (engl. Reinforced learning). Opetuksessa koneelle annetaan palkintoja sekä rangaistuksia riippuen tämän kyvystä onnistua työntäen mallia toivottuun suuntaan. Opetuksen tavoitteena on saada malli oppimaan ympäristöstään tämän tekemien virheiden ja voittojen perusteella. [6]

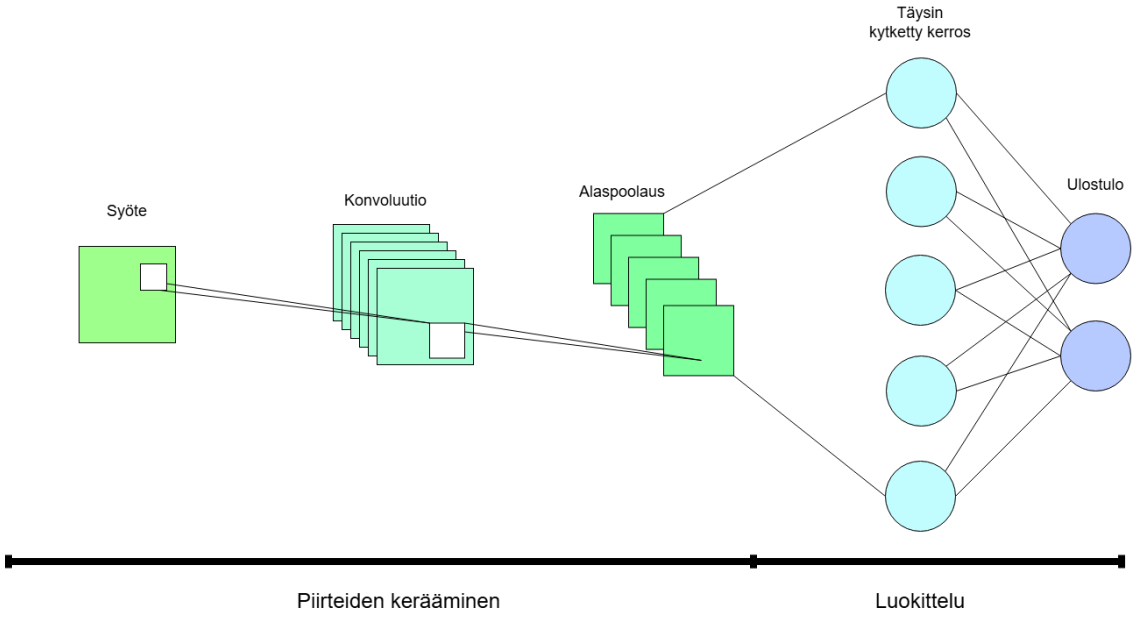
Koneoppimisesta on alaluokka syväoppiminen, joka pystyy yhdistetyn datan esitykseen ja tehtävän oppimiseen. Syväoppimismallin koulutuksessa käytetään raakaa dataa, jonka avulla nämä tunnistavat ja osaavat luokitella piirteitä käyttäen valittua koneoppimismenetelmää. [7] Syväoppiminen pohjautuu keinotekoiseen neuroverkkoon (engl. Artificial neural network, ANN), joka on koneoppimismenetelmä. Keinotekoiset neuroverkot sekä syvät keinotekoiset neuroverkot (engl. Deep ANN) ovat tehokkaita tunnistamaan uhkia sekä poikkeuksia. Verkkojen rakenne koostuu syötekerroksesta, piilokerroksesta ja ulostulokerroksesta. [6]



Kuva 2.2: Esimerkki yksinkertaistetusta neuroverkosta ja syväneuroverkosta, joka pohjautuu artikkelin [6] vastaavaan kuvaan

Toinen tunnettu syväoppimismalli on konvoluutioneuroverkko (engl. Convolutional neural network, CNN). Molemmat mallit sisältävät kerroksia, mutta konvoluutioneuroverkko kykenee vähentämään parametriensä määrää sekä automaattisesti

tunnistaa ympäristön piirteitä hyödyntämällä konvoluutiokerroksia. Konvoluutioverkko hyödyntää konvoluutiokerrosten jälkeen alaspoolaus-kerroksia, joissa karsitaan kerättyjen vektorien ulottuvuuksia. Lopuksi yhtä tai useampaa täysin kytkettyä kerrosta käytetään lopullista luokittelua varten. [6] Konvoluutioverkkoja voidaan hyödyntää varsinkin kuvantunnistuksessa [9]. Syväoppiminen ei ole välttämättä tehokkaampi kuin pinnallisesti koulutettu koneoppimismalli vaan molemmat koneista toimivat paremmin näille suotuisissa tehtävissä [8].



Kuva 2.3: Esimerkki konvoluutioverkosta, joka pohjautuu artikkelin [6] vastaavaan kuvaan

Luonnollisen kielen käsittelyn mallit. Luonnollisen kielen käsittely on tietojenkäsittelytieteen alan aihealue, jonka tavoitteena on saada tietokone tunnistamaan sekä ymmärtämään ihmisen tuottamaa kieltä. Luonnollisen kielen käsittely perustuu kielen rakenteen sekä merkityksen tutkintaan [6]. Menetelmät pohjautuvat tietokoneavustettuihin joukkoon teknologiaa sekä tapoihin analysoida tekstiä. Luonnollisen kielen käsittely koetaan tekoälyn yhdeksi muodoksi [10]. Luonnollisen kielen käsittelyä käytetään valtaosin tiedonkeräämiseen sekä tietokoneen avulla kielen kääntämiseen [6].

Poikkeamantunnistus mallit. Poikkeamantunnistus mallit hyödyntävät koneoppimista analysoidakseen verkkoliikenteessä tapahtuvaa epäilyttävää ja normaalista poikkeavaa toimintaa kuten mahdollisten DDoS hyökkäysten piirteiden tunnistusta [6]. Nämä mallit hyödyntävät tyypillisesti valvomatonta oppimista [11] ja tutkimus osoittaa, että valtaosa yrityksistä, jotka hyödyntävät valvomatonta koneoppimismallien koulutusta käyttävät sitä poikkeaman tunnistus mallien kouluttamiseen [8]. Poikkeaman tunnistus mallit ovat elintärkeitä tunnistamaan mahdollisia poikkeuksia tai pahaenteistä toimintaa järjestelmissä tai verkkoliikenteessä [2].

3 Käyttökohteet

Tässä luvussa tutkimme tarkemmin tekoälyn soveltamisen muotoja kyberturvallisuudessa. Tutkimme kuinka tekoälyä voidaan hyödyntää mahdollisten uhkien tunnistamisessa ja havaitsemisessa kuten myös järjestelmien valvomisessa. Tarkastelemme koneoppimismallien soveltamista verkkoliikenteen valvonnassa kuten myös haittaohjelmien sekä tietojenkalastelun havaitsemisessa. Lisäksi käsittelemme tekoälyn kykyä toimia itsenäisesti keventäen taakkaa työntekijöiltä kriittisiltä kyberturvallisuus sektoreilta. Tarkastelun tavoitteena on hahmottaa miten tekoälyä voidaan hyödyntää kyberturvallisuudessa.

3.1 Uhkien tunnistus

Tutkijat yrittävät jatkuvasti löytää uusia monipuolisia tapoja kehittää turvatoimenpiteitä, joilla voidaan havaita uhkia tarkasti ja automaattisesti. [9] Tunnetuimmat koneoppimisen tietoturva sovellukset ovat verkkohyökkäysten-, haittaohjelmien- ja tietojenkalastelun tunnistus [8]. Taulukossa 3.1 käsitellään aineistoissa esiintyvät keskustelut tekoälyn käytöstä uhkien tunnistamiseen.

3.1.1 Verkkohyökkäysten tunnistus

Tekoälyn avulla voidaan valvoa useita esiintyviä uhkia eri kohteista, joista voidaan hälyttää sekä ennustaa mahdollisia uhkia perustuen vaarantumisindikaattoreihin (engl. indicators of compromise, IOC) [11]. Mahdollisia uhkia voidaan tunnis-

Aineisto	Verkkohyökkäysten tunnistus	Haittaohjelmien tunnistus	Tietojenkalastelun tunnistus
A.S.Kamruzzaman et al (2024) [4]	X		
A.A Siam et al (2025) [2]	X	X	X
N. Ahmed et al (2025) [11]	X		
A. Alsarhan et al (2023) [12]			X
D. L. Antunes ja S. L. Sanchez (2023) [13]			
E. Iturbe et al (2023) [5]			
G. Apruzzese et al (2023) [8]	X	X	X
G. Petihakis et al (2024) [14]			
M. B. Chhetri et al (2024) [15]	X		X
P. Bányász et al (2024) [16]			X
Q. Zeng (2025) [6]	X	X	X
R. A. Hagen et al (2025) [17]	X		
S. R. Ahmed et al (2024) [9]		X	
G. Saranya et al (2025) [18]			X
V. Gudur ja S. K. Ramavath (2026) [1]		X	
W. Badawy (2024) [19]			X

Taulukko 3.1: Aineistojen käsittely uhkien tunnistuksesta

taa verkkoliikenteessä ajoissa käyttäen automaattisia uhkien tunnistus malleja sekä valvomattoman oppimisen malleja ja ennakoivan analytiikan tuella voimme tukea turvaoperaatioita ja ennustaa uusia hyökkäyksiä [11][19]. Verkkohyökkäysten tunnistuksessa etenkin valvomattoman koulutuksen malleja hyödynnetään valvomaan verkkoliikennettä sekä käyttäjien toimintaa ja järjestelmissä olevia lokitietoja, jotka voisivat hälyttää mahdollisista uhista [2].

IDS järjestelmiä hyödynnetään verkkoliikenteen valvontaan ja tämä luokitellaan kahteen luokkaan: NIDS ja HIDS. Koneoppimismalleja on käytetty 2010 luvun alusta alkaen tukemaan verkkopohjaisia tunkeutumisen havaitsemisjärjestelmiä (engl.

network intrusion detection system, NIDS). NIDS malleja asennetaan verkkoympäristöön tunnistamaan mahdollisia turvariskejä monille eri tekijöille kuten älylaitteille tai pilvipalveluille. NIDS hyödyntää koneoppimista tunnistamaan laajan määrän mahdollisia uhkia. [8]

3.1.2 Haittaohjelmien tunnistus

Haittaohjelmien tunnistaminen on yksi kyberturvallisuuteen liittyviä päätehtäviä ja suurimpia haasteita [9][8]. Hakkerit luovat uusia haittaohjelmia päivittäin ja levittävät näitä maailman ympäri valtiosta toiseen kerätäkseen tavallisilta ihmisiltä rahaa [6].

Tyypilliset haittaohjelmien tunnistus tekniikat ovat staattisia sekä riippuvaisia mahdollisten hyökkäyksien piirteistä eivätkä pysty tunnistamaan mahdollisia zero-day tai monimuotoisia hyökkäyksiä. Heuristinen analyysi on todettu olemaan tehokas tapa tunnistaa haitallisen toiminnan piirteitä, mutta tämä antaa usein vääriä positiivisia tai täysin erehtyy ja ei onnistu tunnistamaan haitallisia piirteitä. Tähän koneoppiminen voi tarjota ratkaisun. [9][6] Hyödyntämällä koneoppimisen valvomatonta ja valvottua opettamista koneoppimismallit voivat oppia tunnistamaan epäilyttäviä kuvioita sekä ryhmittää toisiinsa liittyvää toimintaa. Valvotun oppimisen avulla koneet voivat oppia tunnistamaan tunnettuja hyökkäyksiä kouluttamalla malli koulutusmateriaalilla, joka sisältää turvallista sekä haitallista toimintaa datassa. Valvomattoman oppimisen avulla koneet voivat tunnistaa taas tuntemattomia uhkia sekä poikkeuksia. Nämä mallit voivat valvoa verkkoliikennettä tai käyttäjiä tunnistaakseen mahdollisia uhkia turvallisuudelle. [2] Voimme myös opettaa koneita haittaohjelmien profiilien perusteella luokittelemaan haittaohjelmat [6].

Käyttämällä osittain ohjattua oppimista (engl. semi-supervised learning) voidaan tukea uhkien sekä haittaohjelmien tunnistusta sitomalla paljon luokittelematonta dataa yhteen luokitellun datan kanssa. Tämä koneoppimisen opetustyyli

mahdollistaa hankalissa tapauksissa rajatun datan avulla tunnistamaan normaalista poikkeavia tai sinnikkäitä kyberuhkatoimijoita. Tämä toimenpide kehittää haittaohjelmien, poikkeuksien sekä tietojenkalastelu yritysten tunnistamista pienemmällä määrällä luokiteltua koulutusmateriaalia. [2]

3.1.3 Tietojenkalastelun tunnistus

Tietojenkalastus (engl. phishing) on huijaus yritys, jonka tarkoitus on manipuloida uhria. Huijauksen tavoitteena on kerätä kriittistä tietoa uhrilta ja peitellä tämä yritys vakuuttavaksi ja aidoksi viestiksi tai sähköpostiksi. Tietojenkalastus yrityksiltä on hankala suojautua, sillä nämä hyökkäykset perustuvat ihmisten heikkouksien hyväksikäyttöön eivätkä perinteisten kyberturva ratkaisujen ohittamiseen. [20][12]

Tietojenkalastus yritykset kehittyvät jatkuvasti ja perinteiset ratkaisut kuten sääntöpohjainen suodatus ei riitä tunnistamaan ja poistamaan kaikkia huijaus viestejä. Tekoälyä on aloitettu hyödyntämään luonnollisen kielen käsittelyn mallien sekä syväoppimisen muodossa [18]. Koneoppimis ja luonnollisen kielen käsittelyn malleja on jo hyödynnetty huijaus yritysten estämisessä. Koneoppimisen avulla voidaan tunnistaa datassa olevia piirteitä ja kuvioita, joiden perusteella voidaan tehdä ennakoivia toimenpiteitä koneoppimismallin analyysin avulla. Koneoppimismallit voivat opetusmateriaalin avulla tunnistaa epäilyttäviä piirteitä lähettäjän URL-osoitteessa tai luokitella viestin epäilyttäväksi sisällön perusteella kuten lähettäjän nimen tai linkkien pohjalta [12].

Tietojenkalastus yritykset ovat myös kehittyneet pois pelkistä viesteistä ja sähköposteista kokonaisiksi nettisivuiksi. Usein huijausviesteissä olevat linkit vievät näille epäilyttäville nettisivuille, jotka on rakennettu uskottaviksi. Koneoppimisalgoritmit kykenevät tunnistamaan näiden nettisivujen epäilyttävän rakenteen harhaanjohtavasta tekstistä ja valheellisista sisäänkirjautumisen muodoista. [12] Monet petolliset verkkosivut torjutaan estämällä eli asettamalla nämä ”mustalle listalle”. Mustien lis-

tojen tehokkuus heikentyy uhkatekijöiden osatessa vaihtaa toimintaansa nettisivulta toiselle. Koneoppimismallit voidaan kouluttaa valvotun oppimisen avulla tunnistamaan verkkosivujen autenttisuus näiden helposti tunnistettavien piirteiden avulla. [8]

Yksi ensimmäisiä koneoppimisen käyttökohteita oli roskapostin tunnistaminen [8]. Luonnollisen kielen käsittelyn mallit pystyvät tunnistamaan mahdollisia kalasteluja tutkimalla sähköpostien sisältöä, kontekstia ja arvioimalla näiden sisältöjen metadatan. Nämä mallit analysoivat mahdolliset erikoiset sähköpostien lähettäjät sekä tekstissä olevan hoputtavan ja epäilyttävän tekstin. Luonnollisen kielen käsittelyn mallit pystyvät tähän tekstin luokittelulla (engl. text classification). Luokittelemalla suuria määriä teksti muotoista dataa käyttäen malleja kuten TF-IDF (engl. Term Frequency-Inverse Document Frequency) tai BERT (engl. Bidirectional Encoder Representations from Transformers) pystyvät luonnollisen kielen käsittelyn mallit tunnistamaan haitallisia sähköposteja sekä URL osoitteita tai näiden liitteitä kuten dokumentteja. [2]

Lähde [8] kertoo tutkimuksien väittävän tunnistaneensa epäilyttäviä verkkosivuja jopa 99 % tarkkuudella käyttämällä koneoppimismalleja tunnistamaan näiden verkkosivujen URL osoitteista kerättyjen piirteiden avulla. Lähde myös kertoo yhden tutkimuksen näyttävän manuaalisesti estettävien verkkosivujen tunnistuksen olevan vain 9 % kun koneoppimisen avulla voitiin tunnistaa jopa 70 % mahdollisista kalastusyrityksistä. [8] Luonnollisen kielen käsittelyn mallit pystyvät myös semanttiseen analyysiin ja nimettyjen entiteettien tunnistuksen (engl. named entity recognition, NER). Koneet erottavat tekstistä arkaluontoista dataa, joita mahdolliset hyökkääjät voisivat aseistaa. Analysoimalla tekstin kontekstia sekä äänensävyä pystyy mallit tunnistamaan, jos kyseessä on mahdollinen petosyritys. [2]

3.2 Automaattinen toiminta

Tietoturvavalvomo (engl. Security Operations Center, SOC) toimii keskipisteenä yritysten tietoturva toimenpiteissä kuten uhkien tunnistuksessa, analysoinnissa sekä näihin uhkiin vastaamisessa ja puolustuksessa. COVID-19 pandemian jälkeen kyberhyökkäysten määrä on kasvanut ja on alkanut käymään kalliiksi yrityksille suojautua näiltä. Hyökkäysten määrän kasvaessa tietoturvavalvomon työmäärä on kasvanut korreloiden hälytysten määrää. Tietoturvavalvomot saavat hälytyksiä erilaisilta systeemeiltä kuten tunkeutumisen havaitsemisjärjestelmiltä. Hälytysten määrän kasvaessa tietoturvavalvomoissa on alkanut esiintymään hälytysväsymystä (engl. alert fatigue) heikentäen valvomoiden kykyä tehokkaasti erotella yksityiskohtia eri hyökkäyksistä ja vastaamaan vakaviin kyberuhkiin. [15]

Tutkimukset osoittavat tekoälyn automaattisen toiminnan olevan yksi painotetuimpia ominaisuuksia tekoälyllä on kyberturvallisuudessa. Tekoälyn koneoppimismallit, syväoppimismallit, luonnollisen kielen käsittelyn mallit sekä poikkeuksentunnistusmallit on kehitetty muovautumaan alan turvallisuus tarpeisiin. [2]

Syväoppimismallit voivat hyödyntää tekniikoita kuten konvoluutioneuroverkkoja, pitkä-lyhytkestomuistiverkkoja (engl. long short term memory LSTM) ja takaisinkytkettyä neuroverkkoa (engl. recurrent neural network, RNN) joiden avulla koneet voivat tunnistaa automaattisen hahmontunnistuksen avulla esiintyvät haittaohjelmat sekä verkkoon tunkeutumiset [1]. Näiden toimintojen avulla voimme hahmottaa tekoälyn soveltuvuutta itsenäiseen toimintaan. On kuitenkin tärkeää muistaa ettei yksikään uhkien tunnistusmalli ole täydellinen. Tämä tarkoittaa sitä, että vääriä hälytyksiä tulee esiintymään. Tämän takia koneoppimista voidaan hyödyntää automaattisesti filtteröimään sekä priorisoimaan tietyt hälytykset toisten edelle. Tämän lisäksi koneoppimismallit voivat luokitella samankaltaiset hälytykset yhteen, jotta mallit voivat yrittää löytää korreloivia tekijöitä ja tunnistaa turvallisuustehtäviin liittyviä merkittäviä syy-seuraussuhteita. [8]

4 Tekoälyn soveltamisen haasteet

Tekoälyn käyttöä rajoittaa ihmisten oma epävarmuus tekoälyn luotettavuudesta [17]. Koneoppimismallit vaativat jatkuvaa uudelleen kouluttamista sekä ylläpitoa. Syväoppimismallit vaativat paljon resursseja eivätkä ole selitettäviä toiminnaltaan. Luonnollisen kielen käsittelyn mallit eivät pysty tulkitsemaan teknistä kieltä ja usean kielen datatietokannat, joista nämä hakevat tietoa voi heikentää mallin kykyä tulkita viestit oikein. [2]

Viimevuosien aikana tekoälyn käyttö on ollut myös kovassa kasvussa sekä huomiossa. Tekoälyn käyttö on levinnyt kansalaisten käyttöön ja tämän mukana usealle eri alalle ja eri teknologian muotoihin. Tekoälyn käyttö on niin liioiteltua, etteivät yritykset pysty arvioimaan tarkasti kuinka suuren taakan he ovat antaneet tekoälylle suorittaa elintärkeitä tehtäviä. Tämä riippuvaisuus tekoälyn toiminnasta on myös luonut tekoälystä suuren maalitaulun, johon hyökkäykset voivat keskittyä. [14][16] Tässä kappaleessa käymme läpi haasteita, joita tekoäly voi tuoda kyberturvallisuuteen. Tunnistettuja uhkia on esitetty Taulukossa 4.1.

4.1 Datan laatu

Suuri osa tekoälyn tehokkuudesta perustuu datan laatuun. Datan laatu sekä saatavuus määrittää tekoälyn kyvyn oppia tunnistamaan mahdollisia uhkia [6]. Tämän takia tekoälyn työtä vaikeuttaa kasvava tarve salata (engl. encrypt) dataa kuten sähköposteja tai esimerkiksi HTTP verkkoliikennettä valvovat koneoppimismallit kuten

Aineisto	Datan laatu	Väärät positiiviset / negatiiviset	AI:n manipulointi	Yksityisyys	Selitettävyys
A.S.Kamruzzaman et al (2024) [4]				X	
A.A Siam et al (2025) [2]		X	X	X	X
N. Ahmed et al (2025) [11]		X	X	X	
A. Alsarhan et al (2023) [12]		X			
D. L. Antunes ja S. L. Sanchez (2023) [13]		X	X		X
E. Iturbe et al (2023) [5]			X	X	X
G. Apruzzese et al (2023) [8]		X	X	X	X
G. Petihakis et al (2024) [14]			X	X	X
M. B. Chhetri et al (2024) [15]	X	X			X
P. Bányász et al (2024) [16]				X	
Q. Zeng (2025) [6]	X		X	X	X
R. A. Hagen et al (2025) [17]		X	X		X
S. R. Ahmed et al (2024) [9]					
G. Saranya et al (2025) [18]					
V. Gudur ja S. K. Ramavath (2026) [1]	X		X	X	X
W. Badawy (2024) [19]	X		X	X	

Taulukko 4.1: Aineistojen käsittely tekoälyn soveltamisen haasteista

ML-NIDS:n kyky valvoa verkkoliikennettä heikentyy, jos tämä on salattu HTTPS protokollan avulla. [8]

Datan laatu ja tämän saatavuus vaikeuttaa myös koneoppimismallien kouluttamista. Organisaatioiden tarve piilottaa omat tietokantansa vaikeuttaa koneoppimismallien arviointia sekä kouluttamista etenkin, jos haluaisimme sisällyttää mahdollisia malleja kyseisiin yrityksiin käytettäväksi. Mahdolliset saatavilla olevat julkiset mallit sisältävät paljon väärin luokiteltua dataa tai liian pieniä ympäristöjä, joissa kouluttaa sekä arvioida koneoppimismallien toimintaa yleisesti heikentäen näiden

soveltamista. [8] Kyberrikolliset voivat myös myrkyttää tietokantoja, joiden avulla voidaan pilata kokonaisia koneoppimismalleja [6]

4.2 Väärät positiiviset ja väärät negatiiviset

Poikkeuksia tunnistavat algoritmit voivat sopeutua vääriin piirteisiin ja täten vahingossa luoda usein vääriä positiivisia tai vääriä negatiivisia [2]. Tekoälyn yksi suurimpia haasteita kyberturvallisuudessa sekä finanssilaitoksilla on suurten väärien positiivisten ja negatiivisten raporttien määrä, joista moni ovat tyhjästä keksittyjä [15][11]. Kyberturvallisuuden takaamiseen liittyy monesti haasteita väärien hälytysten vuoksi, jotka aiheutuvat tekoälyn toimesta. Turvallisuuden takaajat joutuvat käymään läpi ylimääräisiä hälytyksiä, joka lykkää oikeiden hälytyksien tutkintaa. Tästä syystä tekoälymallien kuten poikkeuksien tunnistusmallien käyttöä rajoitetaan näiden taipumuksesta hälyttää vääristä tilanteista [2][11]. Osa hyökkäyksistä myös keskittyvät tekoäly mallien huijaamiseen. Näiden manipuloivin hyökkäysten tarkoituksena on luoda lisää valheellisia raportteja ja hälytyksiä. [11]

Tekoälyn valheellisten tulosten raportointia voidaan vähentää monella keinolla. Tekoälymalleja voidaan yhdistää keskenään yhdistelmä menetelmillä (engl. ensemble techniques) luoden kokonaisuuksia, jotka ovat tarkempi ja tuottavat tarkempia tuloksia. Malleja voidaan tukea vahvistusoppimisella sekä optimoida tunnistamisen liittyviä kynnyksiä kehittääkseen mallien kykyä tunnistaa vain tiettyjä piirteitä vähentäen yleisistä piirteistä johtuvia vääriä hälytyksiä. [11]

4.3 AI:n manipulointi

Tekoälyn ja tarkemmin koneoppimisen soveltamisen yleistyessä eri aloilla mahdollisiin algoritmiteihin aletaan luottamaan liian paljon ja näistä aletaan olemaan riippuvaisia. Tekoälyn käytön kasvaessa myös hyökkäykset suunnattu kohti käytetty-

jä algoritmeja ovat kasvussa. [13] Tekoälyyn kohdistuvat manipulointi hyökkäykset (engl. adversarial attacks) hyödyntävät tekoälyn heikkoa suojausta ja kykyä tunnistaa näihin kohdistuvia hyökkäyksiä [14]. Hyökkäysten tarkoituksena on manipuloida algoritmeja tuottamaan valheellisia tuloksia tai räätälöimään näiden antamia vastauksia [13]. Manipulointi hyökkäykset ovat väistämättömiä uhkia kyberturvallisuudelle näiden kyvystä hyväksikäyttää koneoppimismallien heikkouksia [6].

Manipulointi hyökkäykset voidaan luokitella myrkytyshyökkäyksiin (engl. Poisoning attack) sekä väistöhyökkäyksiin (engl. Evasion attack). Myrkytyshyökkäykset sijoittuvat koneoppimismallin varhaiseen koulutusvaiheeseen, jossa mallin koulutusdata ”myrkytetään” eli alkuperäinen data pilataan vääristelyllä. Väistöhyökkäysten tarkoituksena on muokata valmiille mallille annettua syötettä, jotta tämä antaisi ei toivottuja vastauksia. [14] Koneoppimisalgoritmeihin hyökätään tyypillisesti kolmella eri tavalla: white-box hyökkäys, black-box hyökkäys ja transfer-hyökkäys. White-box hyökkäyksessä hyökkääjä uskoo tietävänsä kaiken koneoppimismallin rakenteesta opetusdataan asti ja kykenee keskittämään hyökkäyksensä mallin sisäisiin parametreihin asti. Black-box hyökkäys pystyy olemaan vain vuorovaikutuksessa mallille annettujen syöttöjen sekä ulostulojen kanssa. Tämä on koska black-box tekoäly mallit viittaavat tekoäly malleihin, joiden toiminta ei ole selitettävissä tai jotka eivät ole läpinäkyviä [3]. Transfer-hyökkäykset tai ”siirtohyökkäykset” hyödyntää syötettävän datan siirrettävyyttä syöttääkseen mallille haitallista dataa. Hyökkäystä edeltävää tietoa mallista ei ole vaan hyökkäys perustuu oletukseen, että yhdelle mallille räätälöity haitallinen data voidaan yleistää kaikkiin koneoppimismalleihin huijatakseen kaikkia malleja. [13]

Manipulointihyökkäyksiltä yritetään suojautua luomalla koneoppimismalleista sisukkaampia. Kouluttamalla mallit manipulointihyökkäyksiä vastaan lisäämällä koulutus dataan haitallista dataa, josta koneoppimismalli voi oppia vastustamaan aitoa sekä muita datassa olevia hyökkäyksiä vastaan. Neuroverkkoja varten algo-

ritmeja ”tislataan”. Algoritmeja koulutetaan kaksivaiheisessa koulutuksessa. Ensimmäisessä vaiheessa mallille annetaan koulutusmateriaalia, joka ei sisällä tarkkaa luokittelua uhasta vaan uhan todennäköisyys pidetään kevennettynä. Tarkoituksena on saada malli tottumaan epäselvyyksiin, ettei tämä koe häirintää mahdollisista manipulointi hyökkäyksistä. Jälkimmäisessä vaiheessa malli koulutetaan alkuperäisellä ei kevennetyllä koulutusmateriaalilla, jotta tästä tulisi tarkempi. [13][14] Mallia voidaan lopuksi myös tukea syötteen puhdistuksella (engl. Input sanitization) jonka tarkoituksena on kouluttaa malli tottumaan haitallisiin syötteisiin kuten mahdollisiin SQL-injektioihin, jotta tämä pystyy jättämään nämä huomioimatta [13][21].

Tämänhetkiset suojautumiskeinot manipuloivia hyökkäyksiä vastaan pohjautuvat tekoäly malleihin, jotka hyödyntävät valvottua oppimista sekä integroivat monimuotoisia algoritmeja kuten syvät neuroverkot. Esimerkiksi Generatiivisia kilpailevia verkostoja (engl. Generative adversarial networks, GAN) on hyödynnetty mahdollisten uhkientunnistuksessa. Nämä kuitenkin sisältävät tekoälyä, joten itse suojautumiskeinotkin voivat olla mahdollisten manipulointi hyökkäyksien kohteita. [14]

4.4 Yksityisyys

Kyberturvallisuuden pääperiaatteet voidaan jakaa CIA kolmioon. Kolmio sisältää kyberturvallisuuden kolme pääpiirrettä: luottamuksellisuus (confidentiality), eheys (integrity) ja saatavuus (availability). Tekoälyn sisällyttäminen kyberturvallisuuteen heikentää tämän luotettavuutta ja täten tämän käyttö on ristiriidassa kyberturvallisuuden yhtä pääperiaatetta vastaan. [8] Tämän sisältäminen luo myös lukuisia eettisiä kysymyksiä liittyen yksityisyyteen sekä suostumukseen [11]. Tekoälyn toimiessa tämän pitää usein päästä käsiksi erittäin yksityisiin tietokantoihin käsiksi. Mahdollisten koneoppimismallein riippuvuus datasta voi osoittautua suureksi uhaksi järjestöjen yksityisyydelle ja olla epäeettistä toiminnaltaan. [2] Koneoppimismallit ja syväoppimismallit tarvitsevat laajasti erilaisia datamassoja, jotka voivat sisältää

paljon yksityistä tietoa. Datan riippuvuus voi luoda myös yksityisyyteen liittyviä ongelmia kuten datan väärinkäyttöä, tietomurtoja tai luvattomia pääsyjä arkoihin tietoihin. [6]

Tutkimukset ehdottavat ratkaisuiksi tekniikoita kuten differentiaalista yksityisyyttä (engl. differential privacy, DP) ja homomorfista salausta (engl. homomorphic encryption) [6] tai yleisiä käytäntöjä kuten datan jakamista tai toimintakelpoisen datan sääntelyä (engl. actionable data regulations) [8]. Differentiaalinen yksityisyys luo jatkuvaa kohinaa analysoitavan datan päälle rajatussa määrässä mahdollistaen datan lukemisen sekä analysoinnin, mutta estäen yksityiskohtien erottamisen täten varmentaen yksityisyyden ja turvallisuuden [11][6]. Tutkimukset ehdottavat kryptografian käyttöä tukemaan turvallisuutta ja estämään mahdollisia tietovuotoja. Homomorfisen salauksen avulla voimme laskentaa suoraan salatulle datalle ilman, että joutuisimme purkamaan tämän välissä. [6] Lähde [11] kertoo lääketieteellisessä toiminnassa organisaatioiden olevan sallittuja välttää mahdollisia rajoituksia käyttäen esitettyjä ratkaisuja differentiaalista yksityisyyttä sekä hajautettua oppimista.

Yksi esitetty ratkaisu datan puutteelle olisi tämän jakaminen ja datan jakamiseen liittyvien käytäntöjen yleistyminen. Etenkin kyberturvallisuudessa joidenkin tietojen jakaminen voi olla yksinkertaista. Yritykset kuten Sophos tai tietokannat kuten CrimeBB voivat jakaa kymmeniä miljoonia luokiteltuja haittaohjelmia. Hajautetun oppimisen (engl. federated learning) kehittyessä voidaan myös jatkossa esittää anonyymista dataa, joiden avulla kouluttaa koneoppimismalleja. Tämä sallisi yritysten ja organisaatioiden julkaisemaan salattua dataa, joilla voisimme kouluttaa koneoppimismalleja. Toimintakelpoiseen dataan liittyvä lainsäätely voisi myös kehittää datan jaettavuutta. Lakeihin liittyvät säännöt datan hintavasta luokittelusta sekä säännöt liittyen datan keruuseen sekä prosessointiin vaikeuttavat koneoppimismallien kehittämistä. Sääntelyviranomaiset voisivat säätää lakeja taipumaan kyberturvallisuuden tueksi mahdollisen datan keruun sekä jakoon liittyen. [8]

4.5 Selitettävyys

Tekoälyn kehitys on ollut viimevuosien aikana vauhdikasta ja voi jatkossa alkaa toimia laillisten sääntelykehysten ulkopuolella, jonka takia selitettävä tekoäly on elintärkeä kyberturvallisuuden tulevaisuudelle [6]. Tekoälyn selitettävyyden puutteellisuus luo epäluottamusta tekoälyn käyttöä kohden etenkin valtion virastoissa. Selitettävän ja helpommin ymmärrettävän tekoälyn soveltaminen voisi luoda enemmän luottoa tekoälyn toimintaan sekä soveltamiseen myös vähentäen mahdollisia ennakkoluuloja, joita tekoälyn käyttäjillä voi olla. [15]

Epävarmuus tekoälyn käyttöön on varsinkin esillä työympäristöissä, joissa tarvitaan nopeaa toimintaa. Työntekijät eivät pysty luotaamaan täysin automatisoituun ympäristöön, jossa tekoälyn luomat virheet pahentavat tilannetta. Tämä usean väärän positiivisen raportoinnin lisäksi ovat johtavia syitä tekoälyn vähäiseen käyttöön tietoturvaloukkausten hallinnassa. Generoivaa tekoälyä voidaan soveltaa kehittämään tietoturvavalvomojen toiminnan sulavuutta, mutta riskeeraa toiminnan selittämättömyydellä hallusinoivansa mahdollisia hälytyksiä ja tekevänsä tarkistusvirheitä luoden lisää taakkaa. Käyttäjillä voi olla ennakkoluuloja ja kognitiivisia vinoumia liittyen joihinkin tekoäly malleihin kuten black-box tekoäly malleihin näiden epäselvän toiminnan vuoksi. [15] Tämä selitettävyyden puute on myös syy miksi joitain malleja kuten neuroverkkomalleja on hankalaa luoda [8]. Selitettävän tekoälyn vallankumouksen uskotaan luovan tekoälymalleista selkeämpiä sekä luotettavampia ja tutkimusten kuuluisi keskittyä tulevaisuudessa selitettävän tekoälyn luomiseen [11].

5 Pohdintaa

Tämä työ on tuonut esiin tekoälyn laajan käytön kyberturvallisuuden tuessa. Tekoälyllä ja koneoppimismalleilla voidaan tukea eri sektorien heikkouksia ja luoda kyberturvallisuudesta tehokkaampi ja luotettavampi. Havaitsimme tyypillisten tekoälyn käyttökohteiden perustuvan poikkeuksien ja mahdollisten uhkien tunnistamiseen sekä ennustamiseen ja torjumiseen käyttäen moninaisia tekoälyn muotoja. Työn perusteella voimme siis todeta tekoälylle olevan käyttöä kyberturvallisuudessa, mutta tämän käytön mukana tulee myös soveltamiseen liittyviä heikkouksia. Aineistojen perusteella näemme Taulukosta 4.1 tyypillisimpien tekoälyn soveltamisen haasteiden liittyen mallien heikkouteen joutua manipuloiduksi ulkopuolisten toimesta sekä tekoälyn toiminnasta syntyvät yksityisyyden rikkeet ja loukkaukset.

Tutkimuksessa käytetyt aineistot rajattiin tuoreimpiin artikkeleihin kyberturvallisuuden sekä tekoälyn olevan nopeassa kehityksessä minkä vuoksi osa lähteistä voi vanhentua hyvinkin lyhyessä ajassa. Aineistot osoittavat tekoälyn olleen jo vuosia käytössä kuten aineistossa [8] jossa osoitetaan koneoppimisen ensimmäisien käyttötarkoitusten ollessa kyberturvallisuudessa roskapostin havaitsemisessa. Teknologian kehityksen mukana aineistot [19][9][14][5] kertovat tyypillisestä tekoälyn käytöstä poikkeavia ratkaisuja ja ehdottavat omia vaihtelevia lähestymistapoja tekoälyn hyödyntämiseen. Tämä osoittaa tekoälyn soveltamisen olevan vasta alkuvaiheilla ja tekoälyn käytölle löytyy varmasti jatkossa uusia käyttötarkoituksia etenkin generatiivisen tekoälyn olevan kasvussa.

Tämän generatiivisen tekoälyn sekä laajojen kielimallien yleistyminen tuo myös mukanaan uusia haasteita. Tekoäly ei ole saatavilla vain kyberturvan turvaajille vaan myös rikolliset hyödyntävät tekoälyä omaksi eduksi. Tämä on vakavaa etenkin laajojen kielimallien yleistyessä, sillä tämä madaltaa kynnystä osallistua kyberrikoksiin ja antaa tukea kokemattomien rikollisten ensimmäisiin askeliinsa kuten lähteissä [4][16] mainitaankin. Generatiivisen tekoäly luo tietojenkalastelun estämiseen haasteita uskottavien syvävääreennösten (engl. deepfake) sekä äänenkloonaamisen (engl. voice cloning) ansiosta. Laajojen kielimallien kyky tuottaa uskottavaa ihmisen luomaa tekstiä ja kykyä teeskennellä olevansa ihminen näemme, kuinka ihmisiä voidaan manipuloida uskomaan valheellista tietoa. Tämän ansiosta hyökkäyksiä voidaan kohdistaa jatkossa pois perinteisistä kyberturvallisuuden menetelmistä taas hyväksikäyttämään ihmisten heikkoutta langeta mahdollisiin harhoihin kuten uskottaviin kalasteluyrityksiin, joita mitkään kyberturvallisuuden turvatoimet eivät voi estää.

Työn vahvuutena voidaan nähdä tämän ajankohtaisen aiheen laaja tarkastelu sekä tämän ansiosta luodut näkökulmat tekoälyn soveltamisen hyötyihin ja haittoihin. Toisaalta tutkimuksen ollessa laaja katsaus tekoälyn hyödyntämisestä tämä luo myös haasteita tarkkojen aiheiden tutkimisessa ja rajoittaa aiheiden syvällisempää tutkimustyötä. Tekisin tutkielman uudelleen rajaamalla tutkimuskysymykseni koskemaan vain tiettyä uhan tunnistustyyppiä tai koskemaan vain tekoälyn soveltamisen heikkouksiin ratkaisujen löytämistä.

6 Yhteenveto

Tässä työssä tutkittiin kirjallisuuskatsauksen muodossa tekoälyn tyypillisimpiä soveltamisen muotoja kyberturvallisuudessa ja näiden haasteita. Tekoälyn kehittyessä nopeaa vauhtia on tärkeää myös hyödyntää tekoälyä kyberturvallisuuden turvaamiseksi.

Ensimmäinen tutkimuskysymys **TK1** saa vastauksen kirjallisuuskatsauksen luvussa 3, jossa käsittelimme tekoälyn soveltamisesta uhkien tunnistamisessa ja tämän automaattisesta toiminnasta. Voidaan todeta tekoälyn ominaisuudella analysoida suuria datamassoja, valvoa sekä tunnistaa mahdollisia uhkia järjestelmissä ja kyky itsenäiseen toimintaan kehittää kyberturvallisuutta.

Tutkimuksen toiseen kysymykseen **TK2** vastattiin luvussa 4, jossa käsiteltiin tekoälyn soveltamisen haasteita kuten Taulukon 4.1 mukaan yleisimmät haasteet: yksityisyyden rikkominen ja tekoälymallien manipulointi.Voidaan todeta tekoälyn soveltamiseen kuuluvien haasteiden olevan suuri este tekoälyn integroinnille ja täyden luottamuksen ansaitsemiselle.

Työssä tunnistettiin mahdollinen jatkotutkimus tutkimukselle. Tässä tutkielmassa käytiin läpi tekoälyn soveltamisen perinteiset muodot ilman syvempää analysointia tai tutkimustyötä. Jatkotutkimuksessa voidaan sukeltaa syvemmälle aiheeseen ja avata aihealueita tarkemmin. Tutkimuksessa mainittiin perinteisistä tekoälyn menetelmistä poikkeavia soveltamisen muotoja, joita voidaan ajan kuluessa tutkia syvemmin aiheiden kehittyessä.

Yhteenvetona voidaan todeta tekoälylle olevan selkeää käyttöä kyberturvallisuuden takaamisessa. Aineistot osoittivat, että monipuolisia tekoälymalleja voidaan hyödyntää monessa kyberturvallisuustehtävässä sekä ympäristössä näiden ominaisuuksien ansiosta. Tekoälyn soveltamisen haasteet heikentävät tekoälymallien luotettavuutta hidastaen näiden integrointia. Tekoälyn kehittyessä keksimme uusia menetelmiä joiden avulla soveltaa tekoälyä kyberturvallisuuden turvaamisessa.

Lähdeluettelo

[1] V. Gudur ja S. K. Ramavath, "Leveraging Generative AI And Reinforcement Learning For Autonomous Cyber Threat Mitigation", *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*, s. 1–6, 2026. DOI: 10.1109/ICAIC67076.2026.11395822. url: https://ieeexplore.ieee.org/document/11395822.

[2] A. A. Siam, M. M. Hassan ja T. Bhuiyan, "Artificial Intelligence for Cybersecurity: A State of the Art", *2025 IEEE 4th International Conference on AI in Cybersecurity, ICAIC 2025*, 2025. DOI: 10.1109/ICAIC63015.2025.10848980. url: https://ieeexplore.ieee.org/document/10848980.

[3] M. C. Savastano, C. Rizzo, C. Fossataro, D. Bacherini, F. Giansanti, A. Savastano, G. Arcuri, S. Rizzo ja F. Faraldi, "Artificial intelligence in ophthalmology: Progress, challenges, and ethical implications", *Progress in Retinal and Eye Research*, vol. 107, s. 101 374, 2025, ISSN: 1350-9462. DOI: 10.1016/J.PRETEYERES.2025.101374. url: https://www.sciencedirect.com/science/article/pii/S1350946225000473?via%3Dihub.

[4] A. S. Kamruzzaman, K. Thakur ja S. Mahbub, "AI Tools Building Cybercrime & Defenses", *International Conference on Artificial Intelligence, Computer, Data Sciences, and Applications, ACDSA 2024*, 2024. DOI: 10.1109/ACDSA59508.2024.10467401. url: https://ieeexplore.ieee.org/document/10467401.

[5] E. Iturbe, E. Rios, A. Rego ja N. Toledo, "Artificial Intelligence for next generation cybersecurity: The AI4CYBER framework", *ACM International Conference Proceeding Series*, 2023. DOI: 10.1145/3600160.3605051;SUBPAGE:STRING:ABSTRACT;CSUBTYPE:STRING:CONFERENCE. url: https://dl.acm.org/doi/pdf/10.1145/3600160.3605051.

[6] Q. Zeng, "A Comprehensive Review on the Applications of Artificial Intelligence in Cybersecurity", *Proceedings of 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy, CSAIDE 2025*, s. 172–179, 2025. DOI: 10.1145/3729706.3729732;SUBPAGE:STRING:FULL. url: https://dl.acm.org/doi/pdf/10.1145/3729706.3729732.

[7] I. E. Naqa ja M. J. Murphy, "What Are Machine and Deep Learning?", *Machine and Deep Learning in Oncology, Medical Physics and Radiology, Second Edition*, s. 3–15, 2022. DOI: 10.1007/978-3-030-83047-2_1/FIGURES/4. url: https://link.springer.com/chapter/10.1007/978-3-030-83047-2_1.

[8] G. Apruzzese, P. Laskov, E. M. D. Oca, W. Mallouli, L. B. Rapa, A. V. Grammatopoulos ja F. D. Franco, "The Role of Machine Learning in Cybersecurity", *Digital Threats: Research and Practice*, vol. 4, 1 2023, ISSN: 25765337. DOI: 10.1145/3545574;ISSUE:ISSUE:DOI. url: https://dl.acm.org/doi/pdf/10.1145/3545574.

[9] S. R. Ahmed, S. J. Mohamed, M. S. Aljanabi, S. Algburi, D. A. Majeed, N. A. Kurdi, M. Al-Sarem ja J. F. Tawfeq, "A Novel Approach to Malware Detection using Machine Learning and Image Processing", *ACM International Conference Proceeding Series*, vol. 1, s. 298–302, 2024. DOI: 10.1145/3660853.3660931;SUBPAGE:STRING:ABSTRACT;CSUBTYPE:STRING:CONFERENCE. url: https://dl.acm.org/doi/pdf/10.1145/3660853.3660931.

[10] E. Liddy, "Natural language processing", 2001. url: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub.

[11] N. Ahmed, M. E. Hossain, I. S. Hossain, Z. Hossain, M. F. Kabir ja N. Begum, "AI-Driven Cyber Security for Safeguarding Critical Infrastructure and Patient Data", *2nd International Conference on Machine Learning and Autonomous Systems, ICMLAS 2025 - Proceedings*, s. 1485–1492, 2025. DOI: 10.1109/ICMLAS64557.2025.10968807. url: https://ieeexplore.ieee.org/document/10968807.

[12] A. Alsarhan, B. Igried, R. M. B. Saleem, M. Alauthman ja M. Aljaidi, "Enhancing Phishing URL Detection: A Comparative Study of Machine Learning Algorithms", *ACM International Conference Proceeding Series*, vol. 1, 2023. DOI: 10.1145/3625343.3625348;TAXONOMY:TAXONOMY:CONFERENCE-COLLECTIONS;WGROUP:STRING:ACM. url: https://dl.acm.org/doi/pdf/10.1145/3625343.3625348.

[13] D. L. Antunes ja S. L. Sanchez, "The Age of fighting machines: The use of cyber deception for Adversarial Artificial Intelligence in Cyber Defence", *ACM International Conference Proceeding Series*, 2023. DOI: 10.1145/3600160.3605077;ISSUE:ISSUE:DOI. url: https://dl.acm.org/doi/pdf/10.1145/3600160.3605077.

[14] G. Petihakis, A. Farao, A. Gr, P. Bountakas, B. Gr, A. Sabazioti, S. Gr, J. C. Polley, C. Xenakis ja X. Gr, "AIAS: AI-ASsisted cybersecurity platform to defend against adversarial AI attacks", *ACM International Conference Proceeding Series*, vol. 1, 2024. DOI: 10.1145/3664476.3669920. url: https://dl.acm.org/doi/pdf/10.1145/3664476.3669920.

[15] M. B. Chhetri, S. Tariq, R. Singh, F. Jalalvand, C. Paris ja S. Nepal, "Towards Human-AI Teaming to Mitigate Alert Fatigue in Security Operations Cent-

res", *ACM Transactions on Internet Technology*, vol. 24, s. 22, 3 2024, ISSN: 15576051. DOI: 10.1145/3670009. url: https://dl.acm.org/doi/pdf/10.1145/3670009.

[16] P. Bányász, T. Szádeczky ja K. B. Váczi, "The relationship between generative artificial intelligence and cybersecurity", *ACM International Conference Proceeding Series*, vol. 1, s. 209–215, 2024. DOI: 10.1145/3670243.3670781; PAGEGROUP:STRING:PUBLICATION. url: https://dl.acm.org/doi/pdf/10.1145/3670243.3670781.

[17] R. A. Hagen, L. Øverlier ja K. Helkala, "Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues", *Digital Threats: Research and Practice*, vol. 6, 4 2025, ISSN: 25765337. DOI: 10.1145/3759260; ISSUE:ISSUE:DOI. url: https://dl.acm.org/doi/pdf/10.1145/3759260.

[18] G. Saranya, M. R. Y. Priya, K. Kumaran, A. S. Yeshwanth, V. Aswathraj ja M. D. Nagendran, "AI-Powered Phishing Detection: A Data-Driven Cybersecurity Approach", *2025 International Conference on Data Science, Agents and Artificial Intelligence, ICDSAAI 2025*, 2025. DOI: 10.1109/ICDSAAI65575.2025.11011572. url: https://ieeexplore.ieee.org/document/11011572.

[19] W. Badawy, "6G-Enabled IoT Networks Cyber Threat Prevention Using Generative AI", *2024 International Conference on Future Telecommunications and Artificial Intelligence, IC-FTAI 2024 - Proceedings*, 2024. DOI: 10.1109/IC-FTAI62324.2024.10950051. url: https://ieeexplore.ieee.org/document/10950051.

[20] M. Khonji, Y. Iraqi ja A. Jones, "Phishing detection: A literature survey", *IEEE Communications Surveys and Tutorials*, vol. 15, s. 2091–2121, 4 2013,

ISSN: 1553877X. DOI: 10.1109/SURV.2013.032213.00009. url: https://ieeexplore.ieee.org/document/6497928.

[21] L. K. Shar ja H. B. K. Tan, "Predicting SQL injection and cross site scripting vulnerabilities through mining input sanitization patterns", *Information and Software Technology*, vol. 55, s. 1767–1780, 10 2013, ISSN: 0950-5849. DOI: 10.1016/J.INFSOF.2013.04.002. url: https://www.sciencedirect.com/science/article/pii/S0950584913000852?via%3Dihub.