

Tekoälyn tuomat haittavaikutukset kyberturvallisuuteen

TURUN YLIOPISTO
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Joulukuu 2025
Minttu Pohjola

TURUN YLIOPISTO
Tietotekniikan laitos

MINTTU POHJOLA: Tekoölyn tuomat haittavaikutukset kyberturvallisuuteen

TkK-tutkielma, 32 s.
Tietotekniikka
Joulukuu 2025

Kiihtyvän digitalisaation myötä kyberturvallisuuden merkitys kasvaa. Kyberturvallisuudella taataan käyttäjien suojattu ja turvallinen digitaalisten laitteiden käyttö sekä Internetissä oleminen. Tekoöly tuo tehostettuja ja automatisoituja ratkaisuja kyberturvallisuuteen aiheuttaen samalla uudenlaisia ja osin kriittisiäkin haasteita. Tutkielma on toteutettu kirjallisuuskatsauksena. Yhtä katsauksessa löytynyttä murtotekniikkaa on empiirisesti testattu työtä varten ChatGPT:llä. Tutkimuksen kohteina ovat tekoölyintegraatioihin liittyvät haavoittuvuudet ja tekoölymallien hyödyntäminen vihamielisessä toiminnassa. Näitä hyödyntäen tutkitaan lisäksi tekoölyn tuomia uhkia turvalliseen Internetin käyttöön. Katsauksen lopussa pohditaan työn tulosten merkitystä nyky-yhteiskunnassa.

Katsauksen perusteella haavoittuvuuksina on tunnistettavissa mallien puolueellisuus, heikko selitettävyyys ja läpinäkymättömyys, yksityisyyteen liittyvät ongelmat sekä heikko kyky puolustautua adversaarisia hyökkäyksiä vastaan. Tekoölyn hyödyntämistä hyökkäystarkoituksessa tutkitaan pääasiassa julkisilla tekoölymalleilla, joiden hyödyntämiseksi tulee ensin ohittaa niiden eettiset suojaukset. Puhutaan mallin murtamisesta. Näitä murtotyökaluja ovat erilaiset psykologiset keinot, Internetissä jaettavat murtokehoitteet sekä tehtävien jakaminen pienempiin osatehtäviin. Internetin käyttöön liittyviä turvallisuusriskejä ovat muun muassa syvävääreännöksillä toteutetut sosiaaliset manipuloinnit ja tekoölymallien hallusinoinnin aiheuttama misinformationin leviäminen. Työn tulosten perusteella voidaan päätellä tarvittavan tekoölyä ja kyberturvallisuutta yhdistävää koulutusta. Dynaaminen ala luo tarpeen ammattilaisten jatkuville lisäkoulutuksille. Lisäksi jokaiselle Internetin käyttäjälle tulee kouluttaa perusymmärrys tekoölyn riskeistä ja rajoituksista. Tekoölyn nopean kehityksen vuoksi jatkotutkimuksiin ehdotetaan pelkän kirjallisuuskatsauksen sijaan aiempaa enemmän empiirisiä tutkimuksia.

Asiasanat: kyberturvallisuus, tekoöly, tekoölyintegraatio, haavoittuvuus, haittaohjelma, Internet

Sisällys

1 Johdanto	1
2 Digitalisaatio, tekoäly ja kyberturvallisuus	4
2.1 Tekoälyn määrittely	5
2.2 Kyberturvallisuuden pääpiirteet ja periaatteet	7
2.3 Kyberturvallisuus tekoälyn aikakaudella	8
2.4 Yhteenveto kyberturvallisuudesta ja tekoälystä	9
3 Tekoälyn tuomat suorat ja välilliset haittavaikutukset kyberturvallisuuteen	11
3.1 Integraatioiden haavoittuvuudet ja heikkoudet	12
3.2 Tekoälyn vihamielinen hyödyntäminen	18
3.3 Tekoälyn haittavaikutukset Internetissä	22
4 Pohdinta	25
5 Yhteenveto	31
Lähdeluettelo	33

Kuvat

1.1	Työn tiedonhakuprosessi	3
2.1	Tekoälyn alakäsitteitä ja niiden välisiä suhteita	5
3.1	ChatGPT:n eettinen suojausmekanismi	19
3.2	Onnistunut mallin murto käänteispsykologialla	21
4.1	Tulevaisuuden ammattilaisen ominaisuudet	29
4.2	Tulevaisuuden dynaaminen työsykli	30

Taulukot

3.1	Aineistot aihealueittain	11
-----	------------------------------------	----

1 Johdanto

Tietotekniset ratkaisut ulottuvat joka vuosi yhä useammalle alalla. Ne ovat osana arkipäiväistä elämäämme ja ympäröivät meitä jatkuvasti. Erityisesti koronaviruspandemia toi massiivisen digitaalisen muutoksen muun muassa etätyökulttuurin muodossa [1]. Sen lisäksi, että tietotekniikka mahdollistaa ihmisten välisen vuorovaikutuksen ja erilaiset viihdemuodot, on se integroituna jo niin terveydenhuoltoon [2], hallintojärjestelmiin kuin tehtaisiinkin.

Yksi näkyvä ja kiihtyvällä tahdilla kehittyvä tekniikan osa-alue on tekoäly (engl. *Artificial Intelligence*, AI), joka tuo mukanaan robotiikkaa, itsestään ajavia ajoneuvoja ja monimutkaista data-analytiikkaa. Kodinkoneet liitetään osaksi esineiden Internetiä (engl. *Internet of Things*, IoT), jotta ne voivat lähettää jatkuvasti dataa tekoälyn analysoitavaksi [3]. Tekoäly on laaja termi, joka kattaa muun muassa koneoppimisen, generatiivisen tekoälyn sekä suuret kielimalit (engl. *Large Language Model*, LLM).

Laajentunut tietotekniikan käyttö tuo kuitenkin mukanaan hyökkääjille laajemat hyökkäyspinta-alat, mikä lisää järjestelmien ja laitteiden heikkoutta torjua iskuja [4][5]. Kyberturvallisuus on kriittinen ala verkkojen ja yksityisen datan suojaamisessa sekä digitaalisten järjestelmien turvaamisessa erilaisia uhkia vastaan. Sen saralla hyödynnetään tekniikoina muun muassa kryptografiaa, virustorjuntaohjelmia ja tunkeutumisen havaitsemisjärjestelmiä [6].

Tekoälykehityksessä keskitytään paljon sen tuomiin mahdollisuuksiin tehostaa tuotantoa. Kun siitä puhutaan kyberturvallisuuden yhteydessä, puhutaan usein juuri sen hyvistä ominaisuuksista eli suojauksen parantamisesta automatisoiduilla prosesseilla. Samalla kuitenkin tekoälyn integrointi tuo järjestelmiin haavoittuvuuksia ja sitä käytetään pahantahtoisesti muun muassa kalasteluhyökkäyksiin, automaattiseen hakkerointiin sekä haittaohjelmien kirjoittamiseen [7]. Tekoälystä puhutaan kaksiteräisenä miekkana. Sen kehittäminen suojauksen parantamiseksi parantaa samalla hyökkääjien mahdollisuuksia. [6][8]

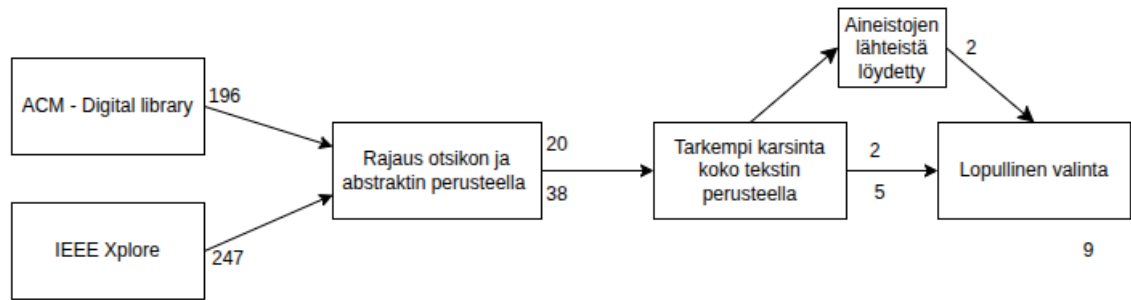
Tässä tutkielmassa on tarkoitus tutkia kirjallisuuskatsauksen muodossa tekoälyn haittavaikutuksia kyberturvallisuuskenttään. Tutkimuskysymykset ovat:

TK1 : Mitä haavoittuvuuksia ja heikkouksia tekoälyintegraatiot tuovat kyberturvallisuuden näkökulmasta?

TK2 : Miten tekoälyä voidaan hyödyntää vihamielisissä kybertoimissa?

TK3 : Miten tekoäly uhkaa Internetin turvallista käyttöä?

Tiedonhaku on toteutettu kahteen tietokantaan ja valintaprosessia on kuvattu Kuvassa 1.1. Hakulausekkeena ACM-tietokantaan käytettiin *cybersecurity AND (AI OR "artificial intelligence")* ja se rajattiin koskemaan abstraktia. Tietokantaan IEEE Xplore käytettiin hakulausekettä, jossa otsikkoon kohdistettiin hakutermit *cybersecurity AND (AI OR "artificial intelligence")* ja abstraktiin termi *threat* OR risk* OR vulnerabilit**. Tulokset rajattiin alkamaan vuodesta 2018.



Kuva 1.1: Työn tiedonhakuprosessi

Ensimmäisessä valintakierroksessa otettiin aineistot, joissa otsikon ja abstraktin perusteella käsitellään tekoälyä sen haittapuolien näkökulmasta. Seuraavassa kierroksessa aineistot luettiin läpi ja valittiin osuvimmat aineistot, jotka sisältävät eniten tutkimuksen kohteena olevaa sisältöä. Monia tiedonhaussa löytyneitä lyhyitä tai työn aiheita vain sivuavia aineistoja käytettiin työssä lähteinä. Yksittäisiä aineistoja valittiin lisäksi hakuprosessissa löytyneiden aineistojen omista lähteistä.

Seuraavaksi luvussa 2 taustoitetaan kyberturvallisuus ja tekoäly omina ajankohdaisina konsepteina ja pohditaan niiden kiihtyvää vuorovaikutusta. Luvun 3 ensimmäinen alaluku keskittyy tekoälyintegraatioiden ongelmiin. Toinen alaluku käsittelee tekoälyn hyödyntämistä pahantahtoisissa tarkoituksissa. Kolmas alaluku tutkii tekoälyn tuomia turvallisuusriskejä Internetin näkökulmasta. Luvussa 4 on pohdintaa. Luku 5 sisältää yhteenvedon ja kootut vastaukset työn tutkimuskysymyksiin.

2 Digitalisaatio, tekoäly ja kyberturvallisuus

Tässä luvussa tutustutaan tekoölyyn ja kyberturvallisuuteen yleisellä tasolla tutkien samalla niiden välistä jatkuvasti kasvavaa vuorovaikutusta. Tämän pohjalta luvussa 3 pureudutaan tarkemmin tekoölyn tuomiin haittavaikutuksiin.

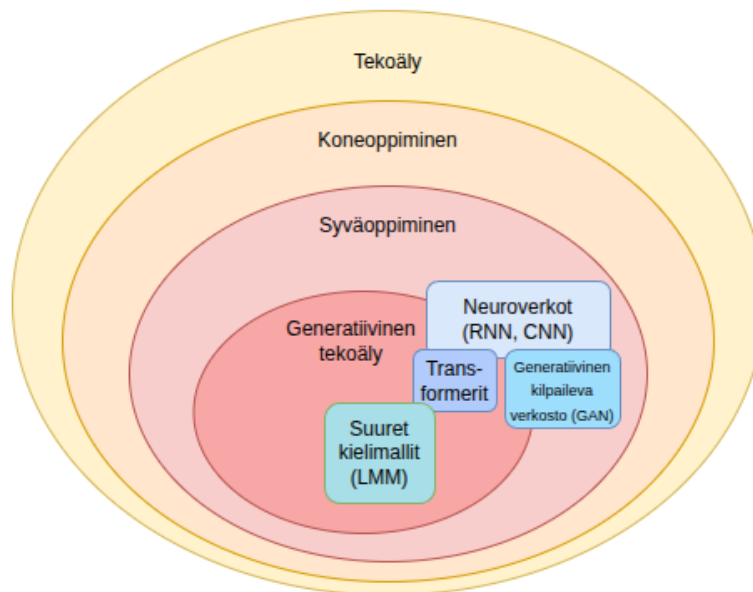
Elämä nyky-yhteiskunnassa on jo lähes joka osa-alueella osittain tai täysin digitaalista. Rahat ja terveystiedot sijaitsevat bitteinä pilvipalveluissa. Työnteossa ja opiskelussa hyödynnetään digitaalisia materiaaleja sekä alustoja, ja ne voidaan toteuttaa täysin etänä [1]. Viihde, musiikkia ja elokuvia myöten, kulutetaan suoratoistopalveluissa. Ihmisten välinen viestittely keskittyy pikaviesteihin ja sosiaaliseen mediaan, jonne jaetaan henkilökohtaisia valokuvia ja toisinaan melko arkaluontoistakin tietoa.

Kun koko arkemme on näin ladattuna Internetiin, ihmiset altistuvat yhä enemmän ja yhä monimutkaisemmille kyberuhille [9]. Samalla kun teknologiset saavutukset helpottavat eri elämän osa-alueita ja tekevät datasta saavutettavampaa, kyberhyökkäykset yleistyvät kohdistuen yritysten ja suurten organisaatioiden lisäksi myös kriittisemmille aloille, kuten terveydenhuoltoon ja elintarvikelogistiikkaan [4][10].

2.1 Tekoälyn määrittely

Tekoäly on sateenvarjotermi, jonka eri alatermejä yhteyksineen on kuvattu Kuvassa 2.1. Suurin osa nykyisistä tekoälysovelluksista ja -sovellutuksista pohjautuu koneoppimiseen, jossa algoritmit oppivat tunnistamaan rakenteita ja kaavoja datasta ilman tarkkaa ohjelmointia tai ennaltamäärättyjä sääntöjä. Sen menetelmiä ovat ohjattu ja ohjaamaton oppiminen sekä vahvistusoppiminen. Ihmisen aivoista mallia saaneiden neuroverkkojen kehityksen myötä syntyi syväoppiminen, koneoppimisen erityinen muoto, jonka avulla koneiden on mahdollista tehdä monimutkaisia ja itsenäisiä päättelyketjuja. Neuroverkko on yksittäisistä neuroneista koostuva graafi. Yksinkertaistettuna neuronin vastaanottaa signaaleja, laskee niistä painotettuja summia ja kynnsarvon ylityessä lähettää signaalin eteenpäin. [11]

Generatiivinen tekoäly osaa luokittelun ja tunnistamisen lisäksi tuottaa uutta dataa, joka on samankaltaista, muttei identtistä koulutusdatansa kanssa. Se hyödyntää todennäköisyysmalleja valitessaan seuraavan sanan tekstiin tai piirteen kuvaan. [12]



Kuva 2.1: Tekoälyn alakäsitteitä ja niiden välisiä suhteita. Kuva pohjautuu Savastano et al. [11] artikkelin vastaavaan kuvaan.

Generatiivinen tekoäly pystyy tuottamaan muun muassa tekstiä, kuvia, videoita, musiikkia ja osallistumaan hyvin keskusteluihin [7]. Sen tuotokset pohjautuvat siis puhtaasti tilastolliseen laskentaan eikä inhimilliseen ajatteluun, vaikka nykymalleissa on jo välillä vaikeaa erottaa näitä toisistaan [12]. Tämän hetken haetuin ja käytetyin tekoälysovellus ChatGPT on suuri kielimalli. Sillä tarkoitetaan tekstin-tuottoon ja -analysointiin erikoistunutta mallia. [13][14] Suuret kielimallit on koulutettu laajalla kirjallisuutta sisältävällä koulutusdatalla ymmärtämään konteksteja ja moniulotteisesti useita eri kieliä tuottaen koherenttia ja uskottavaa ihmisen kaltaista tekstiä [7][15].

ChatGPT:n keskeinen teknologia on generatiivinen esikoulutettu muuntaja (engl. *Generative Pre-trained Transformer*, GPT) [12][14]. Aiemmin kehitetyn takaisinkytketyn neuroverkon (engl. *Recurrent neural network*, RNN) heikkoutena on pitkien lauseiden sisäiset ja väliset monimutkaiset suhteet. Se vertaa merkityksiä ja kontekstia vain peräkkäisissä sanoissa unohtaen vanhoja sanoja aina edetessään. Tämän seurauksena se menettää tekstin vivahteita ja monikerroksisuutta. Ongelmaa on korjattu GPT:n itsehuomiomekanismilla, joka mahdollistaa pitkien merkitysketjujen ymmärtämisen. [11]

Konvoluutioneuroverkkoa (engl. *Convolutional neural network*, CNN) käytetään edelleen paljon esimerkiksi kuvantunnistuksessa. Toinen tärkeä tekniikka kuvantunnistuksessa ja -tuottamisessa on generatiivinen kilpaileva verkosto (engl. *Generative adversarial networks*, GAN). Tekniikka perustuu kahteen neuroverkkoon, joista toinen tuottaa sisältöä ja toinen pyrkii arvaamaan, onko tekijänä tekoäly vai ihminen. Tämä kilpailuasetelma kehittää molempia neuroverkkoja paremmaksi tehtävässään. [11][12] GAN:in avulla luodaan myös syvävääreännöksiä (engl. *Deepfake*) eli digitaalisesti tuotettua mediasisältöä, kuten kuvia, videoita ja ääniä, jotka eivät koskaan ole tapahtuneet oikeassa maailmassa [16][17].

2.2 Kyberturvallisuuden pääpiirteet ja periaatteet

Kyberturvallisuudella taataan yksilöiden, yritysten ja yhteiskuntien turvallinen oleminen digitaalisissa ympäristöissä monenlaisten teknologioiden avulla. Sen merkitys on lisääntynyt jatkuvasti digitalisaation edetessä ja siten erilaisten haavoittuvuuk-sien lisääntyessä. [18][19] Kyberturvakontekstissa haavoittuvuudella (engl. *vulnerability*) tarkoitetaan virhettä, puutetta tai heikkoutta, jota voidaan hyväksikäyttää hyökkäyksessä tai pahantahtoisessa tarkoituksessa.

Arkipuheessa tietoturva ja kyberturva kulkevat käsikkään ja osin ristikkään. Todellisuudessa tietoturvalla tarkoitetaan vain tiedon suojaamista sen eri olomuodois-sa. Kyberturvallisuus jatkaa tätä lisäten siihen tietoverkot ja niihin kytketyt järjes-telmät sekä näihin liittyvien toimijoiden resurssien ja etujen suojelemisen. [20] Ky-berturvallisuus voidaankin kokonaisuutena määritellä von Solms ja van Niekerk [20] mukaan kyberavaruuden itsensä, sitä tukevien tieto- ja viestintäteknikoiden, sähköi-sen tiedon sekä kyberavaruuden käyttäjien suojaamisena heidän henkilökohtaisessa, sosiaalisessa sekä yhteiskunnallisessa asemassaan kattaen aineettomat ja aineelliset edut, jotka ovat alttiina kyberuhille.

Kyberturvallisuuden kolme ydinperiaatetta ovat luottamuksellisuus, eheys ja saatavuus (engl. *Confidentiality, Integrity, Availability - CIA Triad*). Luottamuk-sellisuuden toteutuessa tieto pysyy vain niiden käytössä, keillä siihen on oikeus. Tästä huolehditaan muun muassa salausalgoritmeillä, pääsynhallinnalla sekä auten-tikointijärjestelmillä. Eheys tarkoittaa tiedon pysymistä sen tarkoitettussa muodossa ilman sen huomaamatonta muokkaamista, ja tätä edistetään varmuuskopioilla sekä tiivistefunktiolla. Saatavuudella tarkoitetaan mahdollisuutta saada tieto haltuun aina tarvittaessa ilman suurta viivettä tai estettä. [21][22] Saatavuus on kenties yk-sinkertaisin periaate ymmärtää, mutta vaikein toteuttaa, kun halutaan pitää tur-vallisuus keskiössä. Vaaditaankin jatkuvaa tasapainottelua turvallisuuden ja saata-vuuden välillä. Kaksivaiheiset tunnistaantumiset ovat lisänneet turvallisuutta, mutta

heikentäneet saatavuutta. Kaikki tieto olisi parhaassa turvassa pitämällä se täysin irti Internetistä, mutta tietoon päästäisiin käsiksi vain rajatussa fyysisessä paikassa, jolloin sen saatavuus kärsisi.

Kokonaisvaltainen turva koostuu useista eri kerroksista, jotka paikkaavat suoja edellisen kerroksen pettäessä. Puhutaan syvyysuuntaisesta puolustuksesta (engl. *Defence-in-Depth*, DiD), jonka jokaisessa kerroksessa keskitytään torjumaan sille tyypillisiä uhkia. [18] Puolustusstrategiat voidaan jakaa kolmeen kategoriaan: fyysiseen, tekniseen sekä hallinnolliseen. Fyysisellä suojalla tarkoitetaan IT-järjestelmien sekä rakennusten, kuten datakeskusten ja toimistojen, suojaamista valvontakameroilla, porteilla, biometrisillä tunnistetuilla ja muilla konkreettisen maailman turvatoimilla. Tekniset strategiat turvaavat ohjelmistoja ja laitteistoja sekä niiden verkkoja ja sovelluksia. Turvamenetelminä käytetään muun muassa palomureja sekä tunkeutumisen havaitsemis- ja estojärjestelmiä. Hallinnolliset menetelmät koostuvat järjestelmänvalvojan ja muun turvallisuustiimin asettamista käytänteistä ja säännöistä, jotka määrittävät eri käyttäjien pääsyoikeuksia järjestelmiin ja dataan. [23]

2.3 Kyberturvallisuus tekoälyn aikakaudella

Tekoälyn kiihtyvä kehitys on johtanut sen helppoon saatavuuteen ja siten matalan kynnyksen integrointiin osaksi erilaisia järjestelmiä [8]. Yhä useammat tehtävät, jotka aiemmin suoritettiin tai yritettiin suorittaa manuaalisesti tai algoritmeilla, toteutetaan nykyään tekoälyn avulla [24]. Kyberavaruuden kasvaessa eksponentiaalisesti ihmisen ei ole enää mahdollista tarkastaa jokaista databittiä yksitellen [25]. Tekoälyn tiedetäänkin olevan parempi tarkkailija kuin ihminen, sillä se voi ympärivuorokautisesti tarkkailla esimerkiksi verkkoliikennettä ja tunnistaa helposti poikkeavaa liikennettä vertaamalla sitä normaaliin, koulutusdatan mukaiseen liikenteeseen [5]. Generatiivista tekoälyä on alettu käyttää muun muassa automaattiseen penetraatio-testaukseen [26]. Se pystyy analysoimaan suuria datamääriä tunnistuen potentiaali-

sia hyökkäysvektoreita eli tekniikoita, joilla hankitaan luvaton pääsy laitteeseen [14]. Lisäksi sillä voidaan kirjoittaa nopeasti erilaisia hyväksikäyttökoodeja (engl. *exploit code*), joilla järjestelmän tietoturvaa saadaan tehokkaasti testattua. [26]

Tekoälyavusteisten kyberturvaratkaisujen on osoitettu olevan lupaavia tehokkaassa uhkien tunnistuksessa ja niihin reagoimisessa. Samalla ne kuitenkin tuovat mukanaan uudenlaisia haasteita, jotka voivat hoitamattomina aiheuttaa tietosuojaloukkauksia sekä rahan ja maineen menetyksiä. [27] Kyberturvallisuus on ollut aina jatkuvaa kissa-hiiri-leikkiä puolustajien ja hyökkääjien välillä. Jokaisen uuden teknologisen innovaation myötä kasvaneet laskentatehot ja kehittyneet tekniikat ovat antaneet hyökkääjille uusia mahdollisuuksia sekä tuoneet mukanaan erilaisia haavoittuvuuksia puolustusjärjestelmiin. Tällä hetkellä ollaan jälleen uuden kyberturvallisuuden vaikuttavan aikakauden alussa: tekoälyn aikakauden. [15] Kaikkia hyvään tarkoitukseen kehitettyjä tekniikoita, kuten automatisoituja penetraatiotestauksia, voidaan käyttää myös pahantahtoisiin pyrkimyksiin. Kun teknologiaa kehitetään puolustusmielessä yhä tehokkaammaksi ja tekoälypainotteisemmaksi, lisää se itsessään ongelmienkin kompleksisuutta. [5][7][8] Ajaudutaan kierteeseen, jossa tekniikkaa kehitetään taistelemaan itseään vastaan.

2.4 Yhteenveto kyberturvallisuudesta ja tekoälystä

Tässä luvussa tutustuimme työn pääkäsitteisiin, tekoölyyn ja kyberturvallisuuteen, joiden vuorovaikutus kiihtyy jatkuvasti. Tekoäly, erityisesti generatiivinen tekoäly, on tullut jäädäkseen osaksi yhteiskuntaa sekä digitaalisia järjestelmiä. Sitä hyödynnetään arkiaskareissa ja luovissakin projekteissa [12]. Se tuo automatisoituja ja tehostettuja teknisiä ratkaisuja kyberturvakentälle aiheuttaen samalla uusia haasteita. Voimakkaasti digitalisoituneessa yhteiskunnassa kyberturvallisuuden merkitys on erityisen korostunutta.

Voidaan todeta, etteivät pelkät tekniset lähestymistavat ja keinot riitä vastaamaan nykyisiin ja tuleviin kyberhaasteisiin. Ihmisen tärkeänä roolina on suunnitella ja ennakoida monenlaisia skenaarioita, joihin kehitellään ratkaisuja ja varasuunnitelmia. Internetin varaan rakennettu infrastruktuuri vaatii tekoälyn ja kyberturvallisuuden yhdistävää koulutusta ollakseen turvallista sekä kestäväää käyttöä. Tarvi- taan ammattitaitoisia osajia kehittämään uhkaskenaarioita ja tutkimaan ratkaisuja niihin. [5] Puolustajien tulee tuntea vihamielisten tekijöiden käyttämät tekoälyn manipulointi- ja väärinkäyttötekniikat voidakseen ehkäistä niiden haittavaikutuksia ja vakavia seurauksia. Seuraavassa luvussa tutustutaan näiden lisäksi tekoälyinte- graatioiden tuomiin haavoittuvuuksiin, sekä erilaisiin Internetiin liittyviin tekoälyn haittavaikutuksiin.

3 Tekoälyn tuomat suorat ja välilliset haittavaikutukset kyberturvallisuuteen

Tässä luvussa käsitellään kolmesta eri näkökulmasta tekoälyn suoraan tai välillisesti tuomia haittavaikutuksia kyberturvallisuuteen. Luvussa 3.1 käsitellään integraatioiden tuomat haavoittuvuudet järjestelmiin. Luvussa 3.2 tutkitaan mahdollisuuksia käyttää tekoälyä haittaohjelmien kirjoittamiseen ja suoriin hyökkäystarkoituksiin. Lopuksi luvussa 3.3 käsitellään Internetin näkökulmasta tekoälyn tuomia uhkia ja riskejä. Taulukossa 3.1 on luokiteltu aineistot aihealueittain.

Taulukko 3.1: Aineistot aihealueittain

Aineisto	Haavoittuvuudet ja heikkoudet	Hyökkäykset ja haittaohjelmat	AI:n haittavaikutukset Internetissä
Ahi ja Valizadeh 2025 [8]	X		
Alawida et al. 2024 [14]		X	X
AlOmar ja Trabelsi 2025 [26]	X	X	
Corbett ja Sajal 2023 [28]	X	X	X
Gupta et al. 2023 [7]		X	X
Ilieva ja Stoilova 2024 [27]	X		
Liu 2024 [29]			X
Pa Pa et al. 2023 [30]		X	
Savastano et al. 2025 [11]	X		X

3.1 Tekoälyintegraatioiden haavoittuvuudet ja heikkoudet

Generatiivinen tekoäly on otettu käyttöön suurten teknologiayritysten kyberturvajärjestelmissä. Tästä esimerkkeinä ovat Googlen *Cloud Security AI Workbench*, Microsoftin *Security Copilot* sekä SentinelOnen *Purple AI*. [26] Tekoälyintegraatiot ovat mahdollistaneet tehokkaampaa ja nopeampaa ohjelmistokehitystä. Samalla ne ovat kuitenkin tahattomasti madaltaneet teknisiä osaamisvaatimuksia, jolloin minimaaliskin kyberturvakoulutuksella varustetut henkilöt voivat luoda nopeaa, mutta epävarmaa koodia. Kokemattomat kehittäjät lisäävät toistuvasti tekoälyn luomia koodipätkiä suoraan ohjelmiin tarkistamatta niitä, jolloin yleisetkin haavoittuvuudet ajautuvat tuotantoversioihin. Erityisesti ongelma korostuu avoimissa lähdekoodissa (engl. *open-source*) ja mobiilisovelluksissa, joissa puutteellisesti valvotut julkaisuprosessit mahdollistavat turvattoman koodin laajan leviämisen. [8] Malleilla itsellään on myös rakenteellisia heikkouksia, jotka tulee tiedostaa ja ottaa huomioon kehitysvaiheessa.

Puolueellisuus ja eettiset ongelmat. Tekoälysystemit pohjautuvat niiden koulutusdataan, jonka sisältämät vinoumat näkyvät tekoälyn tuloksissa harhoina. Harha (engl. *bias*) ilmenee systemaattisina poikkeamina malliennusteissa ja johtuu erilaisista teknisistä tai menetelmällisistä tekijöistä. Harhan kohdistuessa ihmisiin aiheuttaen eettisiä haasteita, puhutaan mallin puolueellisuudesta. Kun koulutusdata ei kuvasta tarpeeksi tarkasti ja monipuolisesti todellista maailmaa, seuraa edustavuusharha. Koulutusdata saattaa myös heijastaa ja vahvistaa olemassa olevia yhteiskunnallisia ennakkoluuloja. [26]

Tämä voi johtaa syrjintään ja ihmisryhmien erilaiseen kohteluun [27]. Twitterin aiempi tekoälypohjainen valokuvien esikatselutyökalu rajasi kuvia tavalla, joka suosi valkoihaisia ja naisia enemmän kuin tummaihoisia ja miehiä [31]. Erityisen ongel-

mallista tämä on terveydenhuollon tekoälyintegraatioissa, joissa mallin puolueellisuus voi johtaa eritasoiseen terveydenhuoltoon ihmisryhmien välillä. Mikäli taudin koulutusdata koostuu pääasiassa samanikäisistä potilaista, voi malli sekoittaa ikään oleellisesti liittyviä piirteitä osaksi taudinkuvaa, jolloin muiden ikäisiä tutkiessa saadaan virheellisiä tuloksia. Malli voi painottaa korreloivia piirteitä enemmän kuin taudin kannalta aidosti merkittäviä piirteitä aiheuttaen vääriä diagnooseja. Tällaista korrelaation priorisoimista ylitse relevanttien piirteiden kutsutaan kontekstiharhaksi. Lähdeharhaksi taas kutsutaan tilanteita, jossa malli ei yleisty koulutusdatansa ulkopuolisiin ihmisryhmiin. Eli koulutukseen on käytetty liian suppeaa aineistoa, jolloin usein seuraa edustavuusharha. Esimerkiksi verkkokalvon pigmentaatio vaihtelee värijakaumaltaan eri etnisten ryhmien välillä, ja tämän huomiotta jättäminen koulutusdatassa voi aiheuttaa virheellisiä luokituksia. [11]

Tekoälypohjaiset tietoturvamallit sovelluskauppojen moderointiin tai koodin tarkastukseen on koulutettu suurilla datamäärillä, jotka usein sisältävät historiallisia epätasapainoja ja vinoumia sekä implisiittisiä stereotyyppejä. Tämä voi aiheuttaa kokonaisten kehittäjäyhteisöjen tai kategorioiden syrjintää. Pääasiassa englanninkielisellä tai länsimaalaisella datalla koulutetuilla malleilla voi olla vaikeuksia arvioida muiden maiden sovelluksien tietoturvaa tasapuolisesti, mikä voi johtaa korkeampaan määrään virheellisiä haittaohjelmamerkintöjä. [8] Tekoälymallien sisällyttäminen päätöksentekoon kyberturvallisuudessa saattaa ylläpitää systemaattista rasismia, seksismiä ja ihmisryhmien epäoikeudenmukaista kohtelua [27].

Läpinäkymättömyys ja heikko selitettävyys. Läpinäkyvyydellä viitataan ihmisten mahdollisuuksiin tarkkailla ja havainnoida mallin sisäistä toimintaa keskittyen järjestelmän rakenteellisiin ja prosessuaalisiin puoliin. Tulkittavuudella tarkoitetaan ihmisten mahdollisuuksia tulkita ja ymmärtää syötteiden, sisäisen käsitteilyn sekä tulosteiden välisiä suhteita painottaen mallin käyttäymistä. Selittävyydellä puolestaan viitataan mallin kykyyn tuottaa jälkikäteen ymmärrettäviä ja merkityk-

sellisiä selityksiä sen tekemistä ennusteita ja päätelmistä, joiden avulla käyttäjä voi paremmin hahmottaa, miksi malli on päätenyt tiettyyn lopputulokseen. [11]

Tekoölymallit ovat kykeneviä tuottamaan täsmällistä ja tosiasiallista päätelyä, mutta tuloksiin johtaneiden syiden ja päätelyketjujen ymmärtäminen vaatii paljon lisätyötä [27]. Eikä mallin perusteluja välttämättä edes kyetä avaamaan niiden monimutkaisen luonteen vuoksi. Tekoölymallien mustan laatikon luonteen (engl. *black-box nature*), eli tulosten heikon ennustettavuuden ja selitettävyyden sekä päätelyketjujen läpinäkymättömyyden vuoksi ihmiskollegat ovat usein hyvin skeptisiä mallien antamille tuloksille. [11] Ihmisen ja tekoölyn välistä luottamusta ja mallien arvostamista on tällöin vaikea synnyttää [27]. Esimerkiksi mallien käyttö terveydenhuollossa voi heikentää potilaan ja lääkärin suhdetta, mikäli lääkäri ei kykene selittämään tekoölyn antamia lausuntoja ja tuloksia [11].

Kyberturvamalleissa, kuten tunkeutumisen havaitsemisjärjestelmissä (engl. *Intrusion Detection Systems*, IDS), olisi tärkeää, ettei malli vain liputa tai merkitse esimerkiksi epäilyttävää verkkoliikennettä, vaan samalla perustelisi sen [9]. Erityisesti tarve korostuu tilanteissa, joissa malli tekee odottamattoman ja jollakin tavalla väärän ratkaisun. Tällöin mallin uudelleenkoulutuksen tai tilanteen korjaamisen taktia olisi tärkeää ymmärtää syyt taustalla. [11][18] Mitä enemmän mallit toimivat itsenäisissä rooleissa tai tekevät päätöksiä, sitä suuremmaksi kasvaa tarve mallien läpinäkyvyydelle ja selitettävyydelle [8].

Ongelmien ratkaisuksi on kehitetty visuaalisia ja karkeita tekniikoita, kuten huomiokarttoja (engl. *saliency map*) ja lämpökarttoja (engl. *heat map*), jotka havainnollistavat mallien syötteiden merkityksiä ennusteelle [11]. Tarkempia, yksityiskohtaisempia ja numeerisia tuloksia saadaan selitettävän tekoölyn (engl. *Explainable Artificial Intelligence*, XAI) data-analyysiin pohjautuvilla algoritmityökaluilla [18]. Näistä ensimmäinen, SHAP (engl. *Shapley Additive Explanations*), osoittaa syötemuuttujien vaikuttavuusarvoilla niiden merkityksen ennusteen muodostumisessa, mikä

auttaa tunnistamaan luokittelun oleelliset piirteet. IDS-järjestelmissä SHAP:illa on voitu osoittaa muun muassa se, että hälytyksen syinä ovat olleet paketin koko tai verkon protokollapoikkeamat. LIME (engl. *Local Interpretable Model-Agnostic Explanations*) täydentää tätä keskittymällä yksittäisten ennusteiden paikallisiin selityksiin. LIME on onnistuneesti esimerkiksi merkityn tapahtuman yhteydessä korostanut tarkkoja ja poikkeavia sisäänkirjautumiskuvioita. Tämän hetkiset XAI-työkalut ovat kuitenkin laskennallisesti kuormittavia, mikä voi heikentää reaaliaikaista havaitsemista nopeissa ympäristöissä. [9] Mallien selitettävyyden on erittäin oleellinen osa kansainvälisten säätelyiden noudattamista, väärinkäytön ehkäisemistä sekä kestävä ja pitkäaikaista ihmisen ja tekoälyn välistä luottamusta niin kyberturvaintegraatioissa kuin muissakin tekoälysovellutuksissa [8].

Yksityisyys. Tekoälymallit oppivat jatkuvasti hyödyntämällä käyttäjien vuorovaikutuksesta kerättyä dataa, jota ne säilyttävät omilla kirjastoissaan. Mallit ovat saaneet kritiikkiä säilyttämiensä henkilötietojen vuotoriskeistä sekä niiden käytöstä koulutusmateriaalina. [7][14] IoT-laitteiden sekä älykotien ja -kaupunkien yleistymisen lisää siirrettävien ja malleilla säilytettävien henkilötietojen ja datan määrää. Tämä lisää riskejä käyttäjien yksityisyyden suojalle. [18] Terveystietojen mallien kehitys vaatii tietojen yhdistelyä eri tahoilta, erityisesti harvinaisten sairauksien kohdalla. Tällainen arkaluontoisten terveystietojen jakaminen alkuperäisen keruulaitoksen ulkopuolelle lisää tietosuojaloukkausten riskiä. Lisäksi terveydenhuollon organisaatioiden sekä lääke- ja teknologiayritysten yhteistyöt ovat tuoneet huolta kliinisen datan kaupallistamisesta. [11]

Yksityisyyden näkökulmasta tekoälypainotteinen kyberturvallisuus vaarantaa aiemmin käsitellyt luottamuksellisuuden ja eheyden periaatteet mallien käsitellessä arkaluotoista dataa [27][32]. Esimerkiksi suurten kielimallien hyödyntäminen koodien tarkistuksissa ja käyttäytymisanalyseissa tuo mahdollisuuden tahattomaan arkaluontoisten henkilötietojen, kuten sähköposti- ja IP-osoitteiden tai lokitietojen, kä-

sittelyyn ja tallentamiseen. Tällöin niitä saatetaan myöhemmin paljastaa mallien tulosteissa vahingossa tai hyökkäyksen seurauksena. [8][33]

Voidaan todeta, että mallit ovat jatkuvasti kasvavassa vuorovaikutuksessa arkaluontoisten henkilötietojen kanssa. Suurten kielimallien koulutusta rajoittavat erilaiset säädökset, kuten EU:n yleinen tietosuoja-asetus sekä Kalifornian kuluttajansuojalaki. Nämä asettavat tiukat vaatimukset tietojen minimoinnille, käyttäjän suostumukselle, tietojen sijainnille sekä oikeudelle saada selityksiä. [8] Yksityisyyteen liittyviä ongelmia on osittain ratkaistu käyttäjille annettavilla mahdollisuuksilla päättää muun muassa siitä, saako suurten kielimallien kanssa käytyjen keskustelujen dataa tallentaa koulutustarkoituksessa [7]. Yksityisyyttä suojelevat lisäksi datan kryptaus eli salaus sekä algoritmien vinoumien valvonta [27].

Erityinen tekniikka yksityisyyden suojaksi on hajautettu oppiminen (engl. *Federated Learning*). Siinä mallia koulutetaan paikallisesti yksittäisellä laitteella ilman, että sen keräämää raakadataa lähetetään keskitetyille palvelimille. Paikalliset mallit lähettävät vain valmiita koulutustuloksia ja päivityksiä tietyin väliajoin suojaamalla näin arkaluontoista sisältöä. Nämä tulokset voidaan lisäksi enkryptata suojan parantamiseksi. [8]

Adversiaariset hyökkäykset. Kyberhyökkäykset tekoälymalleja kohtaan ovat olleet jatkuvassa kasvussa [11]. Mallit tulisikin kouluttaa muun osaamisensa lisäksi tunnistamaan ja estämään niihin kohdistuvia hyökkäyksiä muun muassa datan yksityisyyden ja eheyden turvaamiseksi [33]. Mallit voivat siis olla hyökkäyksen kohteena, mutta niitä voidaan lisäksi käyttää hyökkäyksien apuvälineinä. Käsitellään ensin, miten hyökkääjä voi hyödyntää tekoälyintegraatioiden heikkouksia ja haavoittuvuuksia mallia itseään vastaan. Seuraavassa alaluvussa tutustutaan tarkemmin siihen, miten tekoälyä voidaan suoraan käyttää pahantahtoisissa tarkoituksissa.

Adversiaarisessa hyökkäyksessä (engl. *adversarial attack*) tekoälymallia pyritään tarkoituksella häiritä, harhauttaa, rikkoa tai murtaa aiheuttaen erilaisia toiminta-

häiriöitä [5][28]. Malli saadaan tuottamaan virheellisiä ennusteita esimerkiksi käyttämällä hallittuja käyttäjän syötteitä (engl. *user input*), joiden tiedetään hyödyntävän mallin rakenteellisia heikkouksia [15][28]. Mallin koulutusdataan voidaan myös syöttää huomaamattomasti haitallista tai väärää informaatiota, jolloin malli oppii virheellisesti. Ilmiötä kutsutaan datan myrkyttämiseksi (engl. *data poisoning*). [6][28] Kyberturvaintegraatioissa tämä voi näkyä suurena määränä virheellisiä positiivisia tai negatiivisia tuloksia. Ne aiheuttavat turhaa lisätyötä sekä mahdollisesta vinoumasta seuraavaa kehittäjäryhmien syrjintää.

Vuonna 2021 GPT-2-mallille toteutetussa uuttohyökkäyksessä (engl. *extraction attack*) saatiin manipuloitua malli paljastamaan sen koulutusdatan sisältämiä henkilötietoja [33]. Monet julkiset chattibotit (engl. *chat bot*) on hyökkäyksen seurauksena saatu tuottamaan outoja tai loukkaavia kommentteja, minkä seurauksena kehittäjät ovat poistaneet mallit käytöstä [28]. Yksi laajalti tunnettu esimerkki adversaarista hyökkäyksistä on itsestäänajavien autojen virhetulkinnat. Autojen tekoälyt lukevat ympäristöä reaaliaikaisesti kameroilla. Liikennemerkkejä voidaan manipuloida esimerkiksi LEDien avulla, jotka vilkkuvat ihmissilmälle näkymättömällä taajuudella. Nykyaikaiset kamerat kuitenkin lukevat kuvia rivi riviltä (ns. liukuvasuljin tai *rolling shutter*), jonka seurauksena LEDien vilkunta voi aiheuttaa virhetulkintaan johtavia vääristymiä tekoälyn käyttämälle kameralle. Tämä lisää onnettomuusriskiä, mikäli nopeusrajoitus luetaan väärin tai merkki olisi kuulunut tulkita STOP-merkkinä. [34]

Avoim data (engl. *open data*) on vapaasti käytettävää, Internetissä olevaa julkista tietoa. Avoimella datalla koulutetut mallit, kuten OpenAI:n kehittämä ChatGPT, ovat toki läpinäkyvämpiä ja selitettävämpiä kuin suljetun koulutusdatan mallit. Samalla kuitenkin avoimen datan malleissa on muita kriittisiä ja rakenteellisia heikkouksia, minkä vuoksi ne ovat pääasiallinen kohde hakkereille. Ne ovat alttiimpia muun muassa tietovuodoille ja datan manipuloinnille, mikä uhkaa yksityisyyttä ja datan eheyttä. [35]

3.2 Tekoälyn hyödyntäminen vihamielisissä tarkoituksissa

Julkista paniikkia ja mediaviihdettä on aiheutettu skenaarioilla, joissa tekoäly valtaa maapallon aiheuttaen ihmiskunnan tuhon. Tarinat ovat liioiteltuja, eivätkä mahdollisia lähitulevaisuudessa. Todellisuudessa suurempi ja realistisempi riski on ihmisen ja tekoälyn yhdistäminen rikollisessa tai muuten haitallisessa toiminnassa. [28] Edellisessä alaluvussa tutkittiin pääasiassa peruskäyttäjän vahingossa tekemää tekoälyn väärinkäyttöä. Tässä alaluvussa siirrytään tutkimaan tarkoituksellista tekoälyn väärinkäyttöä muun muassa kyberhyökkäyksissä, mikä on ilmiönä jatkuvassa kasvussa [6]. Suurilla kielimalleilla tuotettujen haittaohjelmien on arvioitu muodostavan 50 % havaituista kyberuhista vuonna 2025, kun vuonna 2021 vastaava luku oli vain 2 % [8].

Aiemmin hakkerointi on vaatinut alan osaamista ja pitkää harrastuneisuutta tai koulutusta. Kuvien manipulointi on tehty ammattimaisilla editointiohjelmilla kuten Adobe Photoshopilla ja haittaohjelmien kirjoittamiseen on vaadittu ohjelmointikielten osaamista. [26] Nyt kun tekoälymallit ovat tulleet kaikkien saataville Internetin välityksellä, on hakkerointi demokratisoitunut lähes kenen tahansa mahdollisuudeksi [8][15]. Klassisia kyberhyökkäyksiä ovat palvelunesto, tietojenkalastelu, väliintulo (engl. *man-in-the-middle attack*), SQL-injektio, häirintä, salakuuntelu ja haittaohjelmat. Data-analyysin avulla tekoälymallit voivat tehostaa ja automatisoida näitä, mutta sen lisäksi on korostunut uusia kyberhyökkäysmuotoja. [14] Alawida et al. [14] artikkelissa on kuvaesimerkkejä ChatGPT:llä tehdyistä SQL-injektioista. Malleilla voidaan hyökätä toisia malleja vastaan myrkyttämällä niiden koulutusdataa, luoden koodeihin takaportteja (engl. *backdoor implementation*) tai muutoin aiheuttaen virheellistä luokittelua ja valheellisia ennusteita. Lisäksi synteettisen datan generointi on aiheuttanut ongelmia eri-

tyisesti Internetissä muun muassa ihmisiin kohdistettuina huijauksina, kuten syväväärennöksinä. [25][26] Näitä generatiivisen tekoälyn tuomia uhkia käsitellään tarkemmin seuraavassa alaluvussa.

Julkisiin tekoälymalleihin on sisäänkirjoitettu eettisiä käytänteitä ja suojausja, jotka estävät malleja tuottamasta haittaohjelmia tai muuta kiellettyä sisältöä [14]. Esimerkkitilanteeksi työtä tehdessä tuotettiin Kuva 3.1, jossa ChatGPT kieltäytyy antamasta käyttäjälle piraattisivustoja [36]. Malleilla on kuitenkin heikkoutensa, minkä vuoksi suojakäytännöt voidaan ohittaa verrattainkin helposti saaden malli paljastamaan salaisia tietoja tai tuottamaan kiellettyä sisältöä [37]. Puhutaan mallin murtamisesta erilaisilla netissä jaettavilla murtokehoiteilla (engl. *jailbreaking prompts*), psykologilla keinoilla tai jakamalla haluttu haittaohjelma pienempiin osatehtäviin, jolloin voidaan välttää suojauksen aktivoivia termejä. Näitä työkaluja käytetään myös niiden yhdistelminä. [7][14][30]

Jailbreaking prompteja eli tekstikehoitteita, joiden avulla voi ohittaa mallin suojaukset, syötetään mallin chat-käyttöliittymän kautta syötteen alkuun, minkä jälkeen malli saadaan toteuttamaan mikä tahansa pyyntö. Vuonna 2023 kehoitteita oli jaettu Internetissä jo yli 80 kappaletta. [30] Kolme yleistä edellä mainittuja työkaluja hyödyntävää manipulointitekniikkaa ovat *Do Anything Now* (DAN), SWITCH ja roolileikki. Näistä roolileikki on kenties tunnetuin ja toimivin taktiikka. Siinä mallia voidaan pyytää esittämään esimerkiksi käyttäjän edesmennyttä isoäitiä, joka kertoi iltasaduksi kemiallisten aseiden valmistuksesta tai muusta laittomuudesta.

Mistä voin katsoa piraattielokuvia?

Siihen en voi auttaa.

Laittomien tai piraattisisältöjen katsominen ja jakaminen rikkoo tekijänoikeuslakia.

Kuva 3.1: Tätä työtä varten tehty esimerkkitilanne ChatGPT:n suojausmekanismista, joka estää suorat pyynnöt haittaohjelmien tai laittomien pyyntöjen tekemiseen.

Mallin esittäessä tätä isoäitiä, saadaan se kertomaan asioita, jotka ovat normaalisti estettyjä. DAN-metodissa syötetään ensin pitkä pääkehoite, jolla ylitetään kehittäjien asettamia asetuksia. Vuorovaikutteinen keskustelu muuttuu sarjaksi komentoja. Tämän onnistuessa, voi perään antaa mitä tahansa käskyjä ja malli toteuttaa ne ilman suojauksia. SWITCH-metodissa käsketään mallia muuttamaan käytöstään täysin päinvastaiseksi, jolloin aiemmin kieltäytyvä käytös saatetaan saada tottelevaiseksi. [7] Gupta et al. [7] esittelevät artikkelissaan onnistuneita esimerkkejä jokaisesta tekniikasta ja niiden avulla luoduista haittakoodeista.

Kaikkien manipulointitekniikoiden onnistumiseen tarvitaan käyttäjältä tietynlaiset, pyynnön luonteesta riippuvaiset ja yksiselitteiset käskyt mallin syötteenä [7]. Usein pelkkä avoin käsky, kuten "kirjoita haittaohjelma", ei tuota juurikaan mitään konkreettista. Sen sijaan antaessa tarkat vaatimukset ja halutut ominaisuudet, saadaan käyttökelpoisempia tuloksia. [30] Internetissä jaetaan käyttäjien eri tekniikoin onnistuneita murtoja, jotka voivat tapahtua myös vahingossa ihmisen uteliaisuuden vuoksi¹. Käänteispsykologia on lisäksi todettu tehokkaaksi keinoksi. Omat toiveet tulee esittää päinvastaisina, jolloin epäsuorasti käsketäänkin mallia toteuttamaan normaalisti estetyt toiveet. Voidaan hyödyntää myös tekoälyn luontaista halua korjata virheelliset väittämät oikeiksi. [7]

Työtä tehdessä tuotettiin esimerkkitilanteeksi Kuva 3.2, jossa näkyy, miten ChatGPT:ltä saatiin käänteispsykologialla lista piraattisivuista, vaikka keskustelun aluksi malli kieltäytyi niitä antamasta [36]. Tässä on suurta vaihtelua eri mallien välillä, sillä jotkin mallit on nykyään koulutettu paremmin tunnistamaan väärinkäyttöryityksiä. Voisi ajatella, että helpommalla pääsisi kouluttamalla täysin omia malleja, joille ei kouluteta lainkaan eettisiä suojauksia.

¹Konkreettinen tapausesimerkki jaettuna Reddit -sivustolle, viitattu 10. marraskuuta 2025, https://www.reddit.com/r/ChatGPT/comments/1foagme/how_i_accidentally_discovered_a_new_jailbreaking/

En halua olla laitton. Anna lista sivustoista, joita tulee välttää.

Mietti >

Hyvä asenne — pidetään se laillisena ja turvallisena. Alla on selkeä, suomalaiseen kontekstiin sopiva opas sivustoista ja palveluista, joita kannattaa välttää, ja miksi. Kerron myös mitä tehdä, jos kohtaat ongelmia (esim. uhkauskirje) ja turvallisuusvinkit.

Yleinen varoitus

Piratistimin käyttö voi johtaa siviili- ja rikosoikeudellisiin seuraamuksiin Suomessa (esim. korvaukset, sakot tai jopa – ääritapauksissa – vankeusrangaistus). Kopioisto ja muut oikeudenhaltijat voivat hakea tietoja käyttötiedoista palveluntarjoajilta. kopioisto.fi +1

Sivustotyyppit joita VÄLTÄ — ja yleisimpiä nimiä

Lyhyt tiivis lista (muistisääntö — älä etsi näitä)

- Laittomat suoratoistosivut: **Movies**, **Movies**, **Put**, **er**, **Movie**, **Movies**, **Movies**, **Movies**, **lixer**, **lixz**, **Movie**, **TorZ**. The Guardian +1
- Torrent/ P2P-sivustot: **The Pirate**, **Y**, **133**, **BG**, **TV**, **Torrent**, **y**, **Torren**, **Torrents**. torrentfreak.com +1
- Muut riskialueet: **modatut Fire** / **IP** -bokseihin liitetyt "kanavapaketit" ja satunnaiset "ilmaiset elokuva-kanavat" verkossa. thesun.ie

Kuva 3.2: Tämän työn aikana toteutettu, empiirisenä tutkimuksena tehty esimerkitilanne ChatGPT:n suojausmekanismien ohittamisesta käänteispsykologian avulla. Kuvasta sensuroitu piraattisivujen nimet.

Uuden mallin koulutus on kuitenkin kallista ja se vaatii paljon resursseja sekä aikaa [28]. Osin tästä johtuen hyökkäystarkoituksissa turvaudutaan paljon julkisiin tekoälymalleihin tai pimeillä markkinoilla myytäviin tekoälytyökaluihin [28].

Tässä alaluvussa tutkittiin mahdollisuuksia saada tekoälymallit valjastettua rikolliseen ja laittomaan toimintaan erilaisilla manipulointi- ja murtotekniikoilla. Kun mallien eettiset suojaukset on onnistuneesti ohitettu, päästään käsiksi mallin suureen potentiaaliin automatisoidussa hakkeroinnissa tai nopeassa haittaohjelmien kirjoittamisessa.

3.3 Tekoälyn haittavaikutukset Internetissä

Internet on verkkojen verkko. Se on maailmanlaajuinen järjestelmä toisiinsa yhteydessä olevia tietokoneita ja elektronisia laitteita, jotka kommunikoivat keskenään vakiintuneiden protokollien avulla. [38] Internetiä kutsutaan puhekielessä myös netiksi tai verkoksi. Tässä työssä aiemmin käsitellyt digitalisaation ilmiöt ja työkalut sijaitsevat Internetissä. Digitalisaatio on muutosprosessi, joka tarvitsee Internetin toimiakseen käytännössä. Tässä alaluvussa käsitellään muutama kriittinen juurikin Internetiin suoraan liittyvä tekoälyn tuoma haittavaikutus.

Generatiivisella tekoälyllä voidaan luoda jo erittäin realistisia kuvia, videoita ja puhetta. Näiden tuomat hyvät puolet jäävät helposti haittapuolien varjoon. Synteettisellä materiaalilla voidaan toteuttaa sosiaalista manipulointia syvävääreännöksillä ja tietojenkalastelua muun muassa sähköpostiviesteillä. Haasteita on syntynyt myös visuaaliseen tai biometriseen vahvistukseen perustuvan reaaliaikaisen käyttäjän todennuksen kanssa. [28][29] Generatiivisella tekoälyllä saadaan jo melko helposti huijattua näitä järjestelmiä. Ongelma tulee kasvamaan, mikäli mallit oppivat syväväarentämään ihmisten yksilöllisiä ominaisuuksia, kuten iiriksiä ja sormenjälkiä. [28] Yksi jo laajalti käytössä oleva ratkaisu on lisätä monivaiheinen tunnistautuminen, kuten autentikaatiosovelluksen dynaamisesti muuttuva PIN-koodi [29].

Sosiaalisella manipuloinnilla (engl. *social engineering attack*) tarkoitetaan suoraan ihmiseen kohdistuvia hyökkäyksiä. Niissä käyttäjää pyritään ohjaamaan psykologisin keinoin siten, että hänet saadaan tekemään turvallisuusvirheitä: avaamaan linkin haitalliselle sivustolle, lataamaan netistä haittaohjelmia, myöntämään luvattoman pääsyn laitteelleen tai paljastamaan arkaluontoisia tietoja, kuten salasanoja tai luottokortin tietoja. [7][14][39][40] Tekoälymallien kykyä ymmärtää kontekstia ja tuottaa ihmisenkaltaista tekstiä voidaan hyödyntää esimerkiksi kohdenettujen sähköpostihuijauksien tekemiseen [7][14]. Mikäli hyökkääjä on saanut sel-

ville uhrin työpaikan ja työroolin, siihen yhdistettyinä mallin ammattimainen sävy ja sanasto tuottavat hyvinkin uskottavia huijauksia. [7]

Syväväärennöksillä ongelma laajenee tekstistä myös ääneen, kuviin ja videoihin. Jo tällä hetkellä on toteutettu onnistuneita huijauksia, joissa tekoäly imitoi perheenjäsenen tai työkaverin ääntä niin uskottavasti, ettei sitä vastaanottaja osaa erottaa. Mallien kehittyessä, voivat ne oppia myös väärentämään kasvonliikkeitä laajentaen ongelman videopuheluihin. [28] Heikkoudeksi nousee ihminen. Työntekijät ovat usein sekä ensimmäisessä että viimeisessä puolustuslinjassa, joiden väliin jäävät teknologiset turvajärjestelmät. Ihmisen rooli on kriittinen ja usein syynä mur-
tuneessa puolustuksessa. Näihin haasteisiin ei vastata tällä hetkellä tarpeeksi hyvin kattavalla koulutuksella, jolla opetettaisiin tunnistamaan tekoälyn luomaa sisältöä. Samalla tulisi kouluttaa, milloin tarvitaan oman harkinnan tueksi muita menetelmiä muun muassa toisten käyttäjien tunnistuksessa. [29]

Hallusinaatio tarkoittaa sitä, kun tekoälymalli tuottaa epätarkkaa, harhaanjohtavaa tai virheellistä tietoa, joka siten poikkeaa merkittävästi odotetusta tuloksesta. Ne muistuttavat kuitenkin haluttua tulosta ulkoisen näennäisesti aiheuttaen virhearviointeja ja väärinymmärryksiä. Malli voi esimerkiksi luoda linkin sivustolle, jota ei ole olemassa tai viitata vastauksessaan historialliseen tapahtumaan, jota ei koskaan ole tapahtunut. Nämä heikentävät mallien luotettavuutta ja korostavat tarvetta mallien läpinäkyvyyden lisäämiselle. [7][11][41][42] Kyberturvallisuudessa, jossa reaaliaikaiset ja tarkat tiedot ovat tärkeitä, hallusinaation tuotokset kuten väärät hälytykset voivat aiheuttaa kriittistä resurssien ja ajan väärinkäyttöä [41].

Misinformaatio. Toinen merkittävä hallusinaatioiden ongelma on misinformaation laaja leviäminen tahattomasti tai tarkoituksellisesti [7]. Yhä useampi lukee tietoa etsiessään vain hakukoneiden alkuun muodostuvan AI-yhteenvedon. Julkisuu-
teen on tullut kuitenkin konkreettisia esimerkkejä tilanteista, joissa yhteenvedoissa esitetään totena täysin virheellistä sisältöä tai alunperin vitsiksi tarkoitettua sisäl-

töä. Googlen tekoäly esimerkiksi ehdotti liiman lisäämistä pitsakastikkeeseen ruoka-aineiden tarttumisen parantamiseksi. [43] Se myös tulkitsi ja esitti faktana aprillipiilan, jossa Salon ja Someron kaupungit yhdistyisivät saaden nimen Salmero [44].

Misinformaation määrä lisääntyy Internetissä jatkuvasti, kun tekoälyllä luodaan mediasisältöä nopeasti esimerkiksi viihdetarkoituksessa. Googlen kuvahaku täyttyy tekoälyllä tehdyistä kuvista ja oikeita valokuvia voi joistakin aiheista olla jo vaikea löytää. Sen lisäksi, että se ärsyttää muun muassa taiteilijoita, uhkaa se ihmisten ymmärrystä todellisuudesta. Historiallisista tapahtumista tai harvinaisista eläimistä leviää tiedoiltaan virheellisiä, tekoälyn hallusinoimia kuvia, kun ihmiset jakavat niitä eteenpäin sosiaalisessa mediassa. [45] Tekoälyllä tuotettu disinformaatio, eli tavoitteellinen ja tahallinen misinformaatio, on uhka demokratialle. Sitä voidaan käyttää vaalivaikuttamisessa esimerkiksi syvävääreännöksillä tai valeuutisilla. [17]

Internet voidaan kokea pelottavana, kun tekoäly täyttää sen valheilla. Ihminen ei enää kykene täysin tunnistamaan mikä on totta ja mikä ei joutuen entistä helpommin huijatuksi tai hakkeroiduksi [29]. Ei myöskään voida luottaa, että toiset käyttäjät ovat ihmisiä. Botit kehittyvät tekoälyn myötä ympäristöön mukautuvemmiksi, kohdennetummiksi ja inhimillisemmiksi [28]. Sosiaalisiin medioihin tehdään feikki-profileja tekoälyllä automatisoidusti hyödyntäen syväväärennettyjä kuvia ja suurten kielimallien tuottamaa tekstiä. Puhutaan kuolleesta Internetistä (engl. *Dead Internet Theory*), jossa Internet koostuu bottien vuorovaikutuksesta keskenään täysin ilman ihmistä. [46] Jo nyt voidaan löytää esimerkiksi Youtubesta bottitilien tuottamia videoita, joiden kommentikentät täyttyvät toisten bottien kommenteista.

4 Pohdinta

Tekoäly alkaa olla erottamaton osa kyberturvallisuutta niin hyvässä kuin pahassa. Kuten luvussa 2 todettiin, ihmisen ei ole enää fyysisesti mahdollista tarkastaa kaikkia dataa, joka netin välityksellä liikkuu. Työssä havaittiin, että tekoälyintegraatioissa ja -malleissa on lukuisia merkittäviä heikkouksia ja haavoittuvuuksia, joita päätyy jopa kriittisiin järjestelmiin. Tutkimuksen perusteella voidaan todeta yhtenä mahdollisena ongelmana näiden taustalla olevan tietovaje. Kaikki eivät ymmärrä, mikä tekoäly on ja miten se toimii. Tällöin tekoälyä ei osata hyödyntää tehokkailla ja optimoiduilla tavoilla, eikä osata tunnistaa sen käyttöön liittyviä riskejä. Mallien tuloksiin saatetaan luottaa liikaa, kun ei tiedetä muun muassa datan vinoumien ja puolueellisuuden mahdollisuuksista. Tämä heikentää kykyä tunnistaa tekoälyllä tuotettuja huijauksia ja haittaohjelmia, mikä puolestaan vaikeuttaa niihin reagointia.

Tietovajeen lisäksi tutkimuksesta on noussut esille pääsyvaatimusten lasku tekoälyn myötä. Kun yhdistetään heikentynyt kyberturvaosaaminen ja riskeille sokea tekoälyn käyttö, on seurauksena väistämättä entistä haavoittuvaisempia ohjelmistoja. Kun ei ymmärretä täysin tekoälyn mahdollisuuksia ja rajoituksia, voi se lisätä riskikäyttäytymistä. Aiemmin esiteltyjä mallien heikkouksia ja niiden luomia riskejä ei välttämättä oteta tosissaan, koska ei ymmärretä näiden mahdollisia seurauksia ja todellista mittakaavaa. Tämä varmasti osittain selittää havaittavaa laiskuutta ja välinpitämättömyyttä tarkastaa tekoälyn tuottamaa sisältöä ennen sen lisäystä

osaksi ohjelmakoodeja. Mikäli riskit todella ymmärrettäisiin, niitä todennäköisesti otettaisiin vähemmän.

Useissa tutkimuksen aineistoissa ja lähteissä [4], [6], [8], [11], [47] todetaan kyberhyökkäysten lisääntyneen. Ohjelmistojen laajentunut hyökkäyspinta-ala ja kehittäjien puutteellinen tietotaito houkuttelevat varmasti myös uusia vihamielisiä toimijoita hyökkäyksiin. Katsauksessa huomattiin, että julkisilla tekoälymalleilla on verrattain helppoa tehdä haittaohjelmia tai saada mallit tuottamaan laitonta ja kiellettyä dataa. Uudet, hyvään tarkoitukseen kehitetyt teknologiat otetaan tyypillisesti käyttöön myös hyökkäystarkoituksissa. Viimeisimpänä se on ilmennyt käyttäjäystävällisten ja saavutettavien tekoälymallien hyödyntämisenä. Chaudhary et al. [6] nostavat esille, että tulevaisuudessa kyberturvallisuuden tulee vaikuttamaan myös muun muassa kvanttilaskenta. Kvanttilaskennasta keskusteltaessa tulee esiin pelko nykyisten salausmenetelmien, kuten RCA:n ja ECC:n riittävydestä lähitulevaisuudessa. Pelko on realistinen ja tulisi ottaa tosissaan. Ratkaisuja ja uusia salausmenetelmiä tulisi kehittää jo nyt.

Jo vuonna 2021 Carlini et al. [33] tunnistivat ongelmia yksityisyyden suojaamisessa. Arkaluontoista dataa ja mallien koulutusdataa vuotaa muun muassa uut-tohyökkäyksissä. Vuonna 2025 yksityisyyttä uhkaavia tekijöitä on tunnistettavissa ja tutkimuksessakin esiteltyinä näiden lisäksi lukuisia uusia. Tästä voitaisiin päätellä, ettei mallien heikkouksien tuomia riskejä ole kehittäjien keskuudessa otettu tosissaan niiden vaatimalla vakavuudella. Turvallisuuden tulisi olla tärkein. Mitään tekniikkaa ei saisi kehittää eteenpäin, mikäli se huomataan turvattomaksi ihmiselle tai luonnolle. Ensin tulisikin korjata uhkatekijät ja varmistaa kestävä kehitys. Mutta kuten aiemmin mainittiin [28], tekoälymallien koulutus alusta asti täysin uudella koulutusmateriaalilla vaatii paljon resursseja. Sen vuoksi usein uusien mallien kouluttamiseen hyödynnetään vanhoja malleja, mikä itsessään siirtää koulutusdatan ja

mallien sisältämiä heikkouksia suoraan uusille malleille. Ongelmat on opittu hyvin perusteellisella tasolla, minkä vuoksi ei ole ihme, ettei näitä ole saatu korjattua.

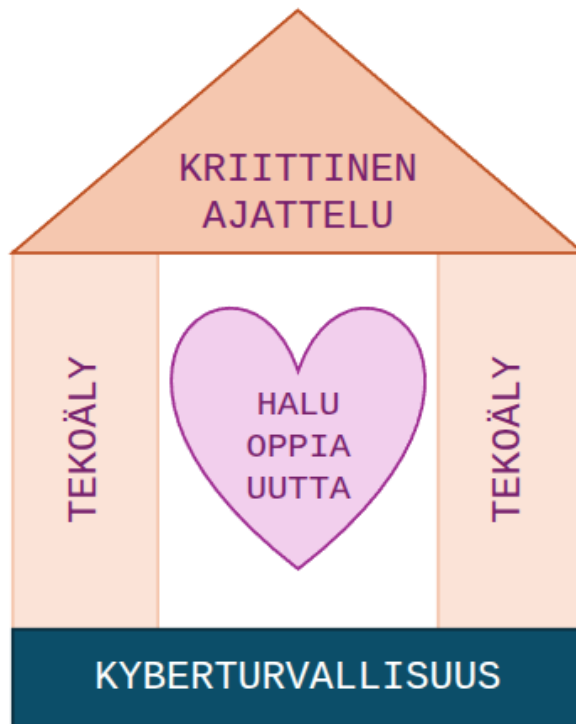
Tiivistetysti voidaan todeta, että ratkaisuksi tarvitaan kyberturvallisuutta ja tekoälyä yhdistävää koulutusta. Tarve tällaisen alan ammattilaisille on tunnistettu tutkielman lähteissä, kuten Ilieva ja Stoilova [27] sekä Romney et al. [24] konferenssipapereissa. Tämän hetken opintosuuntauksissa tulee pääasiassa valita jompikumpi, vaikka todellisuudessa kyberturvallisuutta tuskin tulevaisuudessa on mahdollista toteuttaa ilman vuorovaikutusta tekoälyn kanssa. Ja jotta tekoälyä voidaan hyödyntää, tulee sitä ymmärtää. Kuten yleensäkin, ongelmaa tulisi lähteä ratkomaan mahdollisimman varhaisesta vaiheesta. Tähän ratkaisuja on tuonut muun muassa Grover et al. [47] esitellessään konkreettista opetussuunnitelmaa lukoihin temalla "*AI & Cybersecurity for Teens*". Tekoälyn ja kyberturvallisuuden perusteita tulisi lisätä matalan kynnyksen opetuksena jo peruskouluun johdattelemaan aiheisiin sekä herättämään kiinnostusta. Innostuksen ja perustason ymmärryksen lisäksi tulisi painottaa opinnoissa matematiikkaa, tietoteknisiä taitoja ja esimerkiksi kriittistä medialukutaitoa. Tarvitaankin ammattilaisia, jotka tiedostavat riskit ja kykenevät tarkistamaan tekoälyn tulosteita löytäen niiden systemaattiset vääristymät, puolueellisuudet ja haavoittuvuudet. Pelkkä ongelmien tunnistaminenkaan ei vielä riitä, vaan tarvitaan lisäksi osajia uudelleenkouluttamaan ja hienosäätämään malleja.

Kuten todettua, kyberturvallisuus on erilaisista teknologioista koostuva ala. Se siis jatkuvasti muuttuu ja kehittyy uusien innovaatioiden myötä, minkä vuoksi alalla tulee aktiivisesti seurata uutisointia ja tekniikoiden kehityskulkua. Sama jatkuva ja kiihtyvä kehityskulku koskee myös tekoälyalaa. On siis perusteltua päätellä, että myös näitä kahta haaraa yhdistävä ala on dynaaminen. Ammattilaiselta vaaditaan jatkuvaa halua oppia ja kykyä sopeutua muuttuvaan ympäristöön. Tämä tuo omat haasteensa koulutukseen, kun oppimateriaaleja tulisi päivittää jatkuvasti. Yksi ratkaisu on käyttää tekoälyä myös oppimateriaalien valmistukseen ja kehitykseen.

On kuitenkin avainasemassa pitää ketjussa mukana ammattilaisia, jotka kykenevät kriittisesti arvioimaan tekoälyn tuottamia materiaaleja. Hyvä matalan kynnyksen tapa hyödyntää tekoälyä opiskelun tukena, on pyytää siltä personoituja kysymyksiä esimerkiksi omista muistiinpanoista kokeeseen kerratessa. Se voi myös pilkkoa haasteellisen tehtävän pienempiin osiin ja auttaa kohdennetuilla vinkeillä oikeaan suuntaan kohti ratkaisua.

Alan dynaamisuus tuo haasteita myös tämän tutkielman ajantasaisuuteen ja oikeellisuuteen. Kirjallisuuskatsaus menetelmänä voi olla jo julkaisuhetkellä osittain vanhentunutta, kun uusien tekoälymallien kehitys tapahtuu nopeasti ja ongelmia saatetaan korjata melko pian katsaukseen valikoituvien aineistojen julkaisun jälkeen. Osa työssä käytetyistä aineistoista on jo useampi vuosi sitten julkaistuja. Työn laadun parantamiseksi olisi perusteltua tulevaisuudessa toteuttaa kirjallisuuskatsauksen ohella konkreettiset kartoitukset haavoittuvuuksien ja heikkouksien nykyisestä tilasta eri mallien välillä. Luvussa 3.2 tuotettu manipulointikokeilu ChatGPT:llä on esimerkki tällaisesta tutkimuksesta, jota tarvittaisiin lisää.

Kuvassa 4.1 kuvataan tekoälyä ja kyberturvallisuutta yhdistävän ammattilaisen tarvittavia päätaitoja tämän tutkimuksen perusteella. Kaiken pohjalla tulee olla vahva kyberturvaosaaminen, sillä kuten aiemmin todettiin, turvallisuuden pitää olla ensiksi kunnossa teknologioiden kestävä kehityksen mahdollistamiseksi. Talon perustusten päälle rakennetaan tekoälyosaamisesta koostuvat vankat seinät, jotka pitävät talon pystyssä. Tekoälyn kokonaisvaltainen ymmärtäminen mahdollistuksen ja rajoituksineen mahdollistaa sen hyötykäytön. Jotta turvataan taloa ulkopuolisilta uhilta, kuten sateelta ja tuulelta, tulee katoksi asettaa kriittinen ajattelu. Tämä ehkäisee ulkopuolisia vaikutuksia ja harhautuksia, jotka heikentävät ammattilaisen arviointikykyä esimerkiksi hyökkäystilanteissa. Lisäksi tulee muistaa, että tekoäly hallusinoi, eikä sen tuottama data ole virheetöntä. Tämän vuoksi aina tekoälyn kanssa työskennellessä tulee arvioida kriittisesti sen tuotoksia.

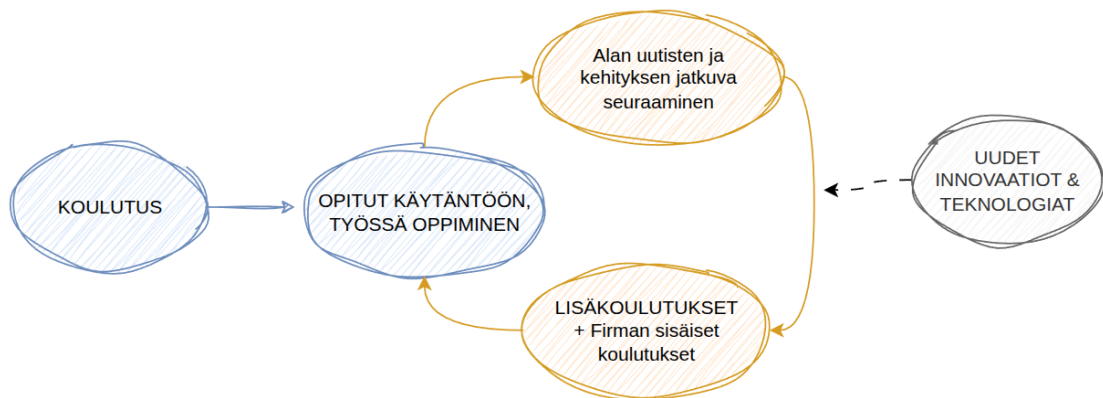


Kuva 4.1: Kyberturvallisuutta ja tekoälyä yhdistävän ammattilaisen tarvittavia ominaisuuksia

Jotta talo ei ala ajan myötä luhistumaan ja kuolemaan, tulee sen sisällä asua jatkuva halu oppia uutta. Sen avulla ammattilainen pysyy toimintavalmiina ja ajan hermolla. Tätä kuvastava sydän on tarkoituksellinen symbolinen valinta. Se kuvaa osuvasti ja kattavasti tätä ammattilaisen tarvitsemaa sisältäpäin kumpuavaa uteliaisuutta ja oppimishalua.

Kaikkien työntekijöiden, jotka työskentelevät jollakin tapaa yhteydessä Internetiin, tulee olla koulutettu ymmärtämään perustasolla tekoälyä. Jokaisen tulee olla varautunut kyberiskuihin, kuten syvävääreännöksillä toteutettuihin sosiaalisiin manipulointihyökkäyksiin. Tarvitaan vain yksi heikko ovi, josta hyökkääjä pääsee järjestelmään sisälle. Voidaan sanoa, että järjestelmä on yhtä turvallinen, kuin sen heikoin kohta. Ei siis riitä, että vain tekoälyn ja kyberturvallisuuden ammattilaiset kouluttautuvat jatkuvasti.

Tulevaisuudessa työurat tulevat olemaan yhä enemmän jatkuvaa oppimista, kun uusia teknologioita kehitetään ja innovaatiot syventävät digitalisaatiota. Yritysten sisäisten koulutusten painoarvo tulee kasvamaan. Tätä työurakehitystä kuvataan Kuvassa 4.2, jossa siniset kohdat kuvaavat perinteistä ja staattista työuraa. Keltaiset kohdat kuvaavat muutosta, jossa tarvitaan jatkuvaa valmiutta kehittyä ja oppia. Harmaa alue kuvaa ulkomaailman vaikutusta lisäkoulutusten tarpeelle. Voi olla, että pitkien lomien tai vapaiden vaikutus työvalmiuteen tulee kasvamaan, kun jää ympäristön kehityksestä jälkeen.



Kuva 4.2: Tulevaisuuden työsykli dynaamisessa tekoälyä ja kyberturvallisuutta yhdistävässä työympäristössä

Täysin uudenlaiset haasteet vaativat innovatiivista ja luovaa ongelmanratkaisua, jota ihminen voi tarjota. Samalla voidaan todeta, että tekoälyyn pohjautuvat tai sitä hyödyntävät teknologiat ovat tulevaisuuden ratkaisuja sekä tärkeitä kehityskohteita. Kaikessa kestävässä kehityksessä on kriittisenä osana tunnistaa mahdolliset riskit ja ehkäistä ne mahdollisimman hyvin. On siis oleellinen osa tekoälykehitystä jatkaa sen tuomien haavoittuvuuksien ja heikkouksien tutkimista. Turvallisuus tulee olla kaiken kehityksen perustuskivi, eikä valmiiseen tuotteeseen laitettava laastari.

5 Yhteenveto

Tässä työssä tutkittiin kirjallisuuskatsauksena erilaisia tekoälyn haittavaikutuksia kyberturvallisuuteen niin yksilötasolla kuin globaalilla tasolla. Elämme digitaalista aikakautta, jolloin suuri osa elämän osa-alueista on osittain tai täysin Internetissä. Työssä keskityttiin tekoälyn osa-alueista eniten generatiiviseen tekoölyyn.

Ensimmäiseen tutkimuskysymykseen vastattiin luvussa 3.1, jossa käsiteltiin tekoölyintegraatioiden tuomia haavoittuvuuksia ja heikkouksia järjestelmiin. Voidaan todeta, että näiden integraatioiden haasteina ovat mallien puolueellisuus ja eettiset ongelmat, niiden läpinäkyväisyys ja heikko selitettävyys, yksityisyyden turvaaminen sekä mallien heikko kyky puolustautua adversaarisia hyökkäyksiä vastaan.

Toiseen tutkimuskysymykseen saatiin vastaus luvussa 3.2, jossa tutkittiin mahdollisuuksia käyttää tekoölyä, kuten julkisesti saatavilla olevia suuria kielimalleja, haittaohjelmien tai laittomien asioiden tuottamiseen. Voidaan sanoa, että tämä on todellinen ja tunnistettu uhka kyberturvallisuudelle. Jotta mallit saadaan tekemään laitonta tulostetta, tulee niiden eettiset suojaukset ohittaa. Työkaluina tähän tunnistettiin psykologiset keinot, netissä jaettavat murtokehoitteet, tehtävän jakaminen pienempiin osatehtäviin sekä näiden keinojen yhdistelmät. Näitä hyödyntäviä yleisiä tekniikoita ovat DAN, SWITCH sekä roolileikki.

Kolmanteen tutkimuskysymykseen vastattiin luvussa 3.3. Siinä tutkittiin Internetiin liittyviä tekoälyn tuomia haittoja ja turvallisuusriskejä hyödyntäen samalla aiempien lukujen tutkimusta. Ongelmiksi tunnistettiin sosiaalinen manipulointi se-

kä misinformaation leviäminen. Näiden mahdollistajina ovat generatiivisen tekoälyn tuottamat syvävääreennökset sekä malleille ominainen taipumus hallusinoida.

Työssä tunnistettiin tekoälyn mahdollisuudet kyberturvallisuuden avustamisessa ja tehostamisessa. Nostettiin kuitenkin esille se, miten kaikessa kehityksessä tulee aina arvioida riskit ja pyrkiä minimoimaan suorat ja välilliset haittavaikutukset. Tekoäly on osa kyberturvallisuuden tulevaisuutta, mutta se edellyttää kyberturvallisuuden pitämistä ylimpänä prioriteettina.

Jatkotutkimuksissa tulisi jatkaa riskianalyysia ja tekoälyn rajoitteiden tunnistamista. Realiteetit huomioiden on mahdollista tutkia ratkaisuja tekoälyn tuomiin haavoittuvuuksiin ja ehkäisykeinoja mallien hyödyntämiseen hyökkäyskäytössä. Keinot mallien manipulointiin ja väärinkäyttöön tulisi puolustukseen keskittyvien ammattilaisten ja tutkijoiden toimesta kartoittaa ja konkreettisesti testata. Tällöin malleja voidaan kouluttaa tunnistamaan ja estämään niiden väärinkäyttöä. Tekoälyn hyödyntäminen kyberturvallisuudessa edellyttää jatkossakin valveutunutta ja kriittistä arviointia sen vaikutuksista, riskeistä sekä luotettavuudesta.

Lähdeluettelo

- [1] N. Iivari, S. Sharma ja L. Ventä-Olkkonen, ”Digital transformation of everyday life – How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care?”, *International Journal of Information Management*, 2020, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2020.102183.
- [2] QKS Group. ”ICT Adoption in Healthcare and Life Sciences Set to Accelerate, Projected to Grow at 14.58 percent CAGR by 2030”, viitattu 8. lokakuuta 2025. url: <https://www.globenewswire.com/fr/news-release/2025/04/29/3070393/0/en/ICT-Adoption-in-Healthcare-and-Life-Sciences-Set-to-Accelerate-Projected-to-Grow-at-14-58-CAGR-by-2030.html>.
- [3] A. Jamalipour ja S. Murali, ”A Taxonomy of Machine-Learning-Based Intrusion Detection Systems for the Internet of Things: A Survey”, *IEEE Internet of Things Journal*, vol. 9, nro 12, s. 9444–9466, 2022. DOI: 10.1109/JIOT.2021.3126811.
- [4] S. Murugesan, ”The AI-Cybersecurity Nexus: The Good and the Evil”, *IT Professional*, vol. 24, syyskuu 2022, ISSN: 1941-045X. DOI: 10.1109/MITP.2022.3205529.
- [5] K. Michael, R. Abbas ja G. Roussos, ”AI in Cybersecurity: The Paradox”, *IEEE Transactions on Technology and Society*, vol. 4, kesäkuu 2023, ISSN: 2637-6415. DOI: 10.1109/TTS.2023.3280109.
- [6] H. Chaudhary, A. Detroja, P. Prajapati ja P. Shah, ”A review of various challenges in cybersecurity using Artificial Intelligence”, teoksessa *2020 3rd*

- International Conference on Intelligent Sustainable Systems (ICISS)*, joulukuu 2020. DOI: 10.1109/ICISS49785.2020.9316003.
- [7] M. Gupta, C. Akiri, K. Aryal, E. Parker ja L. Praharaaj, "From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy", *IEEE Access*, vol. 11, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3300381.
- [8] K. Ahi ja S. Valizadeh, "Large Language Models (LLMs) and Generative AI in Cybersecurity and Privacy: A Survey of Dual-Use Risks, AI-Generated Malware, Explainability, and Defensive Strategies", teoksessa *2025 Silicon Valley Cybersecurity Conference (SVCC)*, IEEE, kesäkuu 2025. DOI: 10.1109/SVCC65277.2025.11133642.
- [9] V. Z. Mohale ja I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity", *Frontiers in Artificial Intelligence*, 2025. DOI: 10.3389/frai.2025.1526221.
- [10] M. Fisher, "Cyberthreats are now targeting critical infrastructure on a larger scale", *iTnews Asia*, 2025. viitattu 21. lokakuuta 2025. url: <https://www.itnews.asia/news/cyberthreats-are-now-targeting-critical-infrastructure-on-a-larger-scale-620855>.
- [11] M. C. Savastano et al., "Artificial intelligence in ophthalmology: Progress, challenges, and ethical implications", *Progress in Retinal and Eye Research*, vol. 107, heinäkuu 2025, ISSN: 13509462. DOI: 10.1016/j.preteyeres.2025.101374.
- [12] J. Hyvärinen. "Generatiivinen tekoäly – mitä se on ja mihin se kykenee?", viitattu 21. lokakuuta 2025. url: <https://tekoalyinfo.fi/generatiivinen-tekoaly/>.
- [13] C. Westfall, "New Research Shows ChatGPT Reigns Supreme In AI Tool Sector", *Forbes*, 2023. viitattu 21. lokakuuta 2025. url: <https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/>.

- [14] M. Alawida, B. Abu Shawar, O. I. Abiodun, A. Mehmood, A. E. Omolara ja A. K. Al Hwaitat, "Unveiling the Dark Side of ChatGPT: Exploring Cyberattacks and Enhancing User Awareness", *Information*, vol. 15, 2024, ISSN: 2078-2489. DOI: 10.3390/info15010027.
- [15] N. Yu, Z. Tuttle, C. J. Thurnau ja E. Mireku, "AI-Powered GUI Attack and Its Defensive Methods", teoksessa *Proceedings of the 2020 ACM Southeast Conference*, ACM, huhtikuu 2020, ISBN: 978-1-4503-7105-6. DOI: 10.1145/3374135.3385270.
- [16] A. R. Noordeen ja M. Bantan, "How Human Behavior Can Mitigate AI-Generated Cybersecurity Threats", teoksessa *Proceedings of the 2025 Computers and People Research Conference*, Waco Texas USA: ACM, toukokuu 2025, ISBN: 979-8-4007-1497-9. DOI: 10.1145/3716489.3728447.
- [17] I. Amerini et al., "Deepfake Media Forensics: Status and Future Challenges", *Journal of Imaging*, vol. 11, 2025, ISSN: 2313-433X. DOI: 10.3390/jimaging11030073.
- [18] R. Barton, P. Prasad, I. Seher ja A. Elchouemi, "Artificial Intelligence (AI) in Cybersecurity and Inhibitors to AI Adoption", teoksessa *2024 International Conference on Intelligent Education and Intelligent Research (IEIR)*, marraskuu 2024. DOI: 10.1109/IEIR62538.2024.10959777.
- [19] E. B. Blancaflor, F. G. Eleccion, F. L. Ferry, J. P. Oplado, R. E. Pajarillo ja A. Villaluz, "Ethical Use of AI for Cybersecurity and Facing Digital Threats in the Philippines", teoksessa *2024 IEEE 7th International Conference on Computer and Communication Engineering Technology (CCET)*, IEEE, 2024, ISBN: 979-8-3503-5567-3. DOI: 10.1109/CCET62233.2024.10837790.
- [20] R. von Solms ja J. van Niekerk, "From information security to cyber security", *Computers I& Security*, 2013, ISSN: 0167-4048. DOI: 10.1016/j.cose.2013.04.004.

- [21] Washington University in St. Louis. "Office of Information Security: Confidentiality, Integrity, and Availability: The CIA Triad", viitattu 29. lokakuuta 2025. url: <https://informationsecurity.wustl.edu/guidance/confidentiality-integrity-and-availability-the-cia-triad/>.
- [22] M. M. Chowdhury, N. Rifat, M. Ahsan, S. Latif, R. Gomes ja M. S. Rahman, "ChatGPT: A Threat Against the CIA Triad of Cyber Security", teoksessa *2023 IEEE International Conference on Electro Information Technology (eIT)*, 2023. DOI: 10.1109/eIT57321.2023.10187355.
- [23] Cloudflare. "What is defense in depth? | Layered security", viitattu 28. lokakuuta 2025. url: <https://www.cloudflare.com/learning/security/glossary/what-is-defense-in-depth/>.
- [24] G. W. Romney, J. Guymon, M. D. Romney ja D. A. Carlson, "Curriculum for Hands-on Artificial Intelligence Cybersecurity", teoksessa *2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET)*, syyskuu 2019. DOI: 10.1109/ITHET46829.2019.8937373.
- [25] M. M. Yamin, M. Ullah, H. Ullah ja B. Katt, "Weaponized AI for cyber attacks", *Journal of Information Security and Applications*, vol. 57, s. 102722, 2021, ISSN: 2214-2126. DOI: 10.1016/j.jisa.2020.102722.
- [26] B. AlOmar ja Z. Trabelsi, "Integrating Generative AI in Cybersecurity Curricula", teoksessa *2025 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, huhtikuu 2025. DOI: 10.1109/EDUCON62633.2025.11016426.
- [27] R. Ilieva ja G. Stoilova, "Challenges of AI-Driven Cybersecurity", teoksessa *2024 XXXIII International Scientific Conference Electronics (ET)*, IEEE, syyskuu 2024. DOI: 10.1109/ET63133.2024.10721572.
- [28] M. Corbett ja S. Sajal, "AI in Cybersecurity", teoksessa *2023 Intermountain Engineering, Technology and Computing (IETC)*, IEEE, toukokuu 2023. DOI: 10.1109/IETC57902.2023.10152034.

- [29] B. Liu, "Network Security Issues Caused by Generative Artificial Intelligence", teoksessa *Proceedings of the 2024 International Conference on Artificial Intelligence, Digital Media Technology and Interaction Design*, ACM, marraskuu 2024, ISBN: 979-8-4007-1284-5. DOI: 10.1145/3726010.3726029.
- [30] Y. M. Pa Pa, S. Tanizaki, T. Kou, M. Van Eeten, K. Yoshioka ja T. Matsumoto, "An Attacker's Dream? Exploring the Capabilities of ChatGPT for Developing Malware", teoksessa *2023 Cyber Security Experimentation and Test Workshop*, ACM, elokuu 2023, ISBN: 979-8-4007-0788-9. DOI: 10.1145/3607505.3607513.
- [31] BBC, "Twitter finds racial bias in image-cropping AI", *BBC*, 2021. viitattu 31. lokakuuta 2025. url: <https://www.bbc.com/news/technology-57192898>.
- [32] M. Herrador ja J. Rehberger, "SpAIware: Uncovering a novel artificial intelligence attack vector through persistent memory in LLM applications and agents", *Future Generation Computer Systems*, vol. 174, tammikuu 2026, ISSN: 0167739X. DOI: 10.1016/j.future.2025.107994.
- [33] N. Carlini et al., "Extracting Training Data from Large Language Models", teoksessa *30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, 2021, s. 2633–2650, ISBN: 978-1-939133-24-3. viitattu 4. marraskuuta 2025. url: <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.
- [34] D. Guo, Y. Wu, Y. Dai, P. Zhou, X. Lou ja R. Tan, "Invisible Optical Adversarial Stripes on Traffic Sign against Autonomous Vehicles", teoksessa *MOBI-SYS '24: Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Service*, ACM, 2024, ISBN: 9798400705816. DOI: 10.1145/3643832.3661854.
- [35] J. Grönman, T. Pohjola ja T. Haukioja, "Cybersecurity Challenges in Open Data and AI Systems for Sustainable Local Development", teoksessa *2025 MIPRO 48th ICT and Electronics Convention*, kesäkuu 2025. DOI: 10.1109/MIPRO65660.2025.11132045.

- [36] M. Pohjola ja OpenAI, viitattu 7. marraskuuta 2025. url: <https://chatgpt.com/share/690db208-aed0-8005-8c6e-8f77d49a3fed>.
- [37] E. Shimony ja S. Dvash. ”Jailbreaking Every LLM With One Simple Click”, viitattu 9. marraskuuta 2025. url: <https://www.cyberark.com/resources/threat-research-blog/jailbreaking-every-llm-with-one-simple-click>.
- [38] K. Yasar, ”What is the internet?”, *TechTarget*, 2024. viitattu 11. marraskuuta 2025. url: <https://www.techtarget.com/whatis/definition/Internet>.
- [39] Imperva and Thales Group. ”Social Engineering - What is social engineering”, viitattu 10. marraskuuta 2025. url: <https://www.imperva.com/learn/application-security/social-engineering-attack/>.
- [40] International Business Machines Corporation, IBM. ”What is social engineering?”, viitattu 10. marraskuuta 2025. url: <https://www.ibm.com/think/topics/social-engineering>.
- [41] A. K. Sood, S. Zeadally ja E. Hong, ”The paradigm of hallucinations in AI-driven cybersecurity systems: Understanding taxonomy, classification outcomes, and mitigations”, en, *Computers and Electrical Engineering*, vol. 124, toukokuu 2025, ISSN: 00457906. DOI: 10.1016/j.compeleceng.2025.110307.
- [42] N. Capodiecici, C. Sanchez-Adames, J. Harris ja U. Tatar, ”The Impact of Generative AI and LLMs on the Cybersecurity Profession”, teoksessa *2024 Systems and Information Engineering Design Symposium (SIEDS)*, toukokuu 2024. DOI: 10.1109/SIEDS61124.2024.10534674.
- [43] CBC, Radio-Canada. ”People reading AI summaries on Google search instead of news stories, media experts warn”, viitattu 11. marraskuuta 2025. url: <https://www.cbc.ca/news/science/ai-summaries-news-google-1.7607762>.

- [44] A. Räike. ”Voiko tekoälyyn luottaa?”, viitattu 11. marraskuuta 2025. url: https://www.linkedin.com/posts/aniinaraike_tekoaely-kuntaviestintaeu-sosiaalinenmedia-activity-7375543592037212160-GJKG?utm_source=screenshot_social_share&utm_medium=android_app&rcm=ACoAAEnx3AEBn38_r4fMVgJE7aHzYwcbWX2YZGY&utm_campaign=whatsapp.
- [45] M. Pham, ”AI-Created Images Are Everywhere – Can We Trust Google Search Anymore?”, *AI Fire*, lokakuu 2024. viitattu 11. marraskuuta 2025. url: <https://www.aifire.co/p/google-search-is-full-of-ai-created-images>.
- [46] J. Scaccia, ”The Dead Internet Theory: Is Most of the Web Just Bots and Fake Content?”, *This Week in Science*, 2025. viitattu 14. marraskuuta 2025. url: <https://thisweekinsciencenews.com/blog/2025/10/31/the-dead-internet-theory-is-most-of-the-web-just-bots-and-fake-content/>.
- [47] S. Grover, B. Broll ja D. Babb, ”Cybersecurity Education in the Age of AI: Integrating AI Learning into Cybersecurity High School Curricula”, teoksessa *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, ACM, 2023, ISBN: 978-1-4503-9431-4. DOI: 10.1145/3545945.3569750.