Topi Luukkanen

VÄLTTÄMISKÄYTTÄYTYMINEN KONEOPPIMISMALLIEN KEHITTÄMISESSÄ

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2020

TIIVISTELMÄ

Luukkanen, Topi
Välttämiskäyttäytyminen koneoppimismallien kehittämisessä
Jyväskylä: Jyväskylän yliopisto, 2020, 68 s.
Tietojärjestelmätiede, Pro Gradu -tutkielma
Ohjaajat: Lehto, Martti & Äyrämö, Sami

Tämä Pro Gradu -tutkielma käsittelee välttämiskäyttäytymistä koneoppimismallien kehitystehtävissä. Tavoitteena oli kasvattaa ymmärrystä suorituskykykriittisissä työympäristöissä työskentelevien käyttäjien välttämismotivaatioon vaikuttavista tekijöistä Teknologiauhkien välttämisen teorian avulla. Tämä tutkielma vaikuttaa olevan teorian ensimmäinen koneoppimisen kontekstiin sijoittuva tutkimus, ja myös ensimmäinen, joka pyrkii mallintamaan käytetyn teorian muuttujia formatiivisesti. Empiirinen tutkimus toteutettiin strukturoidun kyselytutkimuksen avulla, johon vastasi 16 koneoppimismalleja työkseen kehittävää käyttäjää, ja kerätty aineisto analysoitiin kvantitatiivisin menetelmin. Varsinaiseen tutkimuskysymykseen tutkielma ei kyennyt tuottamaan luotettavaa vastausta, mutta teorian ja koneoppimisen tutkimuksen kehityksen kannalta tutkielma tuotti huomionarvoisia tutkimustuloksia. Indikaattorien formatiivinen mallinnustapa mahdollistaa muuttujan konseptuaalisen kentän kuvaamisen kattavasti, mikä on tärkeää, kun muuttuja toimii osana välttämiskäyttäytymisen kaltaista kompleksista prosessia. Ollessaan aihepiirinsä ensimmäinen tutkimus, tämä tutkielma toimii myös lähtöpisteenä käyttäjälähtöisen koneoppimisen tutkimukselle sekä esittää Teknologiauhkien välttämisen teorialle uuden formatiivista mallinnustapaa hyödyntävän tutkimushaaran.

Asiasanat: koneoppiminen, formatiivinen mallintaminen, TTAT, välttämiskäyttäytyminen

ABSTRACT

Luukkanen, Topi
Avoidance Behavior in Machine Learning Model Development
Jyväskylä: University of Jyväskylä, 2020, 68 pp.
Information Systems Science, Master's Thesis
Supervisors: Lehto, Martti & Äyrämö, Sami

This Master's Thesis studies user avoidance behavior in the context of Machine Learning model development. The goal was to increase understanding about the factors affecting avoidance motivation of users working in performance critical environments with the help of Technology Threat Avoidance Theory. This thesis appears to be the first reserach of the theory to attempt modeling the latent factors formatively. Empirical research was conducted as a structured questionnaire, to which 16 answers were collected from users developing Machine Learning models as a part of their work. The data collected was analyzed with quantitative methods. The thesis failed to provide an answer to the initial research question, but from the perspective of the theory and Machine Learning research, meaningful findings were produced. Formative variable modelling approach allows the latent variable's conceptual domain to be covered comprehensively, which is important, when the variables of interest act as a part of a complex process such as user avoidance behavior. Being the first study of its kind, this thesis offers a starting point to user-focused Machine Learning research and presents the Technology Threat Avoidance Theory a new branch that utilizes formative modelling approach.

Keywords: machine learning, formative modeling, TTAT, avoidance behavior

KUVIOT

Kuvio 1 Negatiivisen ja positiivisen palautesilmukan erot (suomennettu Liang & Xue, 2009, s. 76 kuviosta) 11
Kuvio 2 Teknologiauhkien välttämisen teorian (TTAT) varianssiteoreettinen näkemys (suomennettu ja muokattu Liang & Xue, 2009, s. 79) 15
Kuvio 3 Ali- ja ylisovittaminen kaksiulotteisessa piirreavaruudessa (Müller, Mika, Rätsch, Tsuda & Schölkopf, 2001, 182 kuviosta) 20
Kuvio 4 Koneoppimisprosessi ja koneoppimisen hyökkäyspinta (muokattu Papernot ym., 2016b, s. 5 kuviosta) 22
Kuvio 5 Esimerkki hyökkäyksestä online-luokittimeen (suom. Barreno ym., 2006, s. 7 kuviosta). Vasemmalla lähtötilanne ja oikealla hyökkäyksen aiheuttama päätöspinnan siirtyminen. 24
Kuvio 6 Tutkielman malli Boysenia ym. 2019, s. 102 mukaillen 33

TAULUKOT

Taulukko 1 Koneoppimiseen kohdistettujen hyökkäysten luokittelu (suom. Barreno ym., 2006, s. 3) 24
Taulukko 2 Puolustustekniikoita koneoppimiseen kohdistettuihin hyökkäyksiin (suom. Barreno ym., 2006, s. 4 kuviosta) 26
Taulukko 3 Kyselytutkimus 31
Taulukko 4 VIF-arvot alkuperäiselle sekä karsitulle mallille 39
Taulukko 5 Indikaattorien latautuminen latentteihin muuttujiin ja PLS-regression painoarvot 40
Taulukko 6 Polkuanalyysin tulokset ja tuet hypoteeseille 42

SISÄLLYS

TIIVISTELMÄ
ABSTRACT
KUVIOT
TAULUKOT

1 JOHDANTO 7
1.1 Keskeiset käsitteet 8

2 TEKNOLOGIAUHKIEN VÄLTTÄMISEN TEORIA 10
2.1 Tausta 10
2.2 Kognitiiviset prosessit 12
2.2.1 Uhkan arviointi 12
2.2.2 Hallintakeinojen arviointi 13
2.2.3 Tilanteen hallinta 13
2.3 Sosiaalinen ympäristö 14
2.4 Teknologiauhkien välttämisen teorian kehitys 15

3 KONEOPPIMINEN 17
3.1 Oppiminen prosessina 17
3.2 Menetelmät 18
3.3 Koneoppimismallin yleistämiskyky 19
3.4 Vihamielinen koneoppiminen 20
3.4.1 Koneoppimisen hyökkäysala 21
3.4.2 Hyökkäysten taksonomia 22
3.5 Hyökkäyksiltä puolustautuminen 25
3.6 Kirjallisuuskatsauksen yhteenveto 27

4 EMPIIRINEN TUTKIMUS 29
4.1 Tutkimusongelma ja hypoteesit 29
4.2 Kyselytutkimus 30
4.3 TTAT-malli 32
4.4 Indikaattorien reflektiivinen ja formatiivinen mallinnus 33

5 ANALYYSI 35
5.1 Datan käsittely 35
5.2 Formatiivistesti mallinnettujen muuttujien analysointi 35

6 TULOKSET 38
6.1 Multikollineaarisuus 38
6.2 Absoluuttiset ja suhteelliset vaikutukset 39
6.3 Polkumuuttujat ja tutkimuksen hypoteesit 41

7 POHDINTA 43
7.1 Tutkimuksen tulokset 43
7.2 Rajoitteet (Limitations) 45
7.3 Kontribuutiot ja jatkotutkimus 45
7.4 Johtopäätökset 46

LÄHTEET 49

LIITTEET 55

1 JOHDANTO

Koneoppimisella toteutettu datan mallinnus on mullistanut useita toimialoja. Muun muassa terveydenhuollon, turvallisuusalan ja taloushallinnon toimintatavat ovat muuttuneet merkittävästi koneoppimismenetelmien tuottaman, päätöksentekoa tukevan, informaation avulla (Papernot, McDaniel, Sinha & Wellman, 2016c). Viime vuosikymmenen aikana koneoppiminen on popularisoitunut ja tekoälyteknologioiden saatavuus on siten laajentunut suljetuista organisaatioiden sisäisistä alustoista yksilöiden ulottuville erilaisten alustapalveluiden, rajapintojen ja ohjelmointikirjastojen, kuten TensorFlow'n AWS SageMakerin, IBM Watsonin, muodossa (Abadi ym. 2016b, Amazon Web Services; IBM Watson). Edistykset koneoppimisen ja muiden tekoälyteknologioiden saralla ovat vastanneet Big Datan synnyttämiin haasteisiin suurten datamassojen prosessoimisesta ja tarjonneet ratkaisuita dataintensiivisiin ongelmiin (Sapp, 2018; Fraley & Cannady, 2017). Jopa kansallisten kriittisen infrastruktuurin suojaamisen keskiössä olevien IDS- ja IPS-järjestelmien (engl. *Intrusion Detection/Prevention System*) lupaavimpia kehityssuuntia hyökkäysten havaitsemisen ja ehkäisemisen tehostamiseksi sekä automatisoimiseksi on löydetty juuri koneoppimisen alle lukeutuvista menetelmistä (mm. Corona, Giacinto & Roli, 2013; Fraley & Cannady, 2017). Koneoppimisen käyttömahdollisuudet eivät kuitenkaan rajoitu pelkkään puolustukseen, vaan käyttäjä voi myös konkreettisesti hyödyntää samoja menetelmiä ja algoritmeja kyberhyökkäysten valmistelun ja suorittamisen välineenä.

Viime vuosikymmenen aikana julkaistut lukuisat tieteelliset lehtiartikkelit ja konferenssijulkaisut ovat osoittaneet, että koneoppimista voidaan hyödyntää vihamielisiin tarkoituksiin jopa erittäin vähäisellä ennakkotiedolla kohdejärjestelmästä (mm. Barreno, Nelson, Joseph & Tygar, 2010; Szegedy ym., 2014). Tästä huolimatta vain hyvin harvat kyberuhkaraportit tunnistavat koneoppimiseen liittyviä uhkia. Tekoälyteknologioiden potentiaalia huomioidaan hyvin kapeakatseisesti vain positiivisessa valossa (mm. Aon 2019; CyberEdge Group, 2019), mikä indikoi, ettei ymmärrys koneoppimisen uhkista ja tekniikoista niiden välttämiseksi saa päivänvaloa tieteellisen yhteisön ulkopuolella. Vihamielisen

koneoppimisen uhkia vastaan kuitenkin kehitetään aktiivisesti mallien resilienssiä kasvattavia tekniikoita sekä hyökkäyksien riskiä ja negatiivisia vaikutuksia minimoivia menetelmiä (mm. Papernot ym., 2016c; Liu ym., 2018).

Välttäytymiskäyttäytymisellä tarkoitetaan henkilön pyrkimystä objektiivisesti vähentää negatiivisen lopputuleman toteutumisen todennäköisyyttä (Liang & Xue, 2009). Viime vuosikymmenen aikana ymmärrys käyttäjien välttämiskäyttäytymisestä teknologiauhkien suhteen on laajentunut valtavasti. Teknologiauhkien välttämisen teoriaa (*Technology Threat Avoidance Theory, TTAT*) (Liang & Xue, 2009) on kehitetty tutkimalla muun muassa käyttäjien salasanaturvallisuutta sekä haittaohjelmien ja vakoiluohjelmien tuomien uhkia välttämistä. Toistaiseksi on kuitenkin vielä tutkimatta, miten käyttäjät pyrkivät välttämään uhkia suorituskykykriittisissä työympäristöissä, kuten koneoppimismallien kehitystöissä, jossa heidän valintansa vaikuttavat myös organisaation turvallisuuteen. Koneoppiminen tutkimuskohteena vaikuttaa olevan TTAT:lle täysin uusi tutkielman kirjoitusajankohdan aikaan. Tämä Pro Gradu -tutkielma pyrkii siten paikkaamaan edellä mainitun tutkimusaukon tutkimalla käyttäjien välttämiskäyttäytymistä koneoppimismallien kehityksessä. Tämän tutkielman tutkimuskysymyksenä on:

- *Mitkä tekijät vaikuttavat käyttäjän välttämiskäyttäytymiseen suorituskykykriittisessä työympäristössä?*

Empiirinen aineisto kerätään strukturoidun kyselylomakkeen muodossa. Aineiston analyysissä hypoteesien testaamiseksi hyödynnetään kvantitatiivista tutkimusotetta. Kirjallisuuskatsauksen tukena hyödynnetään tutkijan omaa kandidaatintutkielmaa. Tutkielman taustatavoitteena on kasvattaa ymmärrystä koneoppimiseen liittyvistä uhkista sekä kehitysyhteisöissä vallitsevista näkemyksistä.

1.1 Keskeiset käsitteet

Tätä tutkielmaa varten laadittu käsite **suorituskykykriittinen työympäristö** määritellään sellaisena ympäristönä, jossa työntekijän työpanoksen arvo määrittyy pitkälti käyttäjän työnjäljen systemaattisen laadun, tarkkuuden, tehokkuuden tai nopeuden mukaan.

ISO/IEC 27000:2018 standardin mukaisesti **tietoturvallisuus** (engl. i*nformation security*) määritellään niinä järjestelyinä, joiden avulla taataan tiedon **luottamuksellisuus** (engl. *confidentiality*), **eheys** (engl. *integrity*) ja **saatavuus** (engl. *availability*). Tietoturvallisuus kuvataan prosessina, joka mahdollistaa liiketoiminnan jatkuvuuden ja yrityksen menestymisen. Sanastokeskus TSK:n (2018) mukaan **tietoturvauhkat** käsittävät tietoturvaan kohdistetut, potentiaaliset tapahtumat tai kehityskulut, jotka toteutuessaan vaarantavat yhden tai useamman tietoturvan periaatteen toteutumisen. Termi **haavoittuvuus** viittaa tietoturvaan kohdistuvaan uhkaan ja alttiuteen sen toteutumiselle. **Resilienssillä** kuvataan valmiutta kohdata kriisitilanteita ja kykyä mukauttaa toimintatapojaan ylläpi-

tääkseen toimintakykynsä odottamattomien turvallisuutta vaarantavien olosuhteiden vallitessa (Sanastokeskus TSK, 2018). **Hyökkäysalalla** (Lehto ym., 2017) (engl. *attack surface*) tarkoitetaan kaikkia niitä tapoja, joiden avulla hyökkääjä voi tunkeutua järjestelmään tai aiheuttaa siihen vahinkoa (Manadhata, & Wing, 2011).

Koneoppimisen aihealueella on hyvin vähän vakiintuneita suomenkielisiä termejä, joten käännöksissä on hyödynnetty tilastotieteellisten termien lisäksi kirjoittajan omaa harkintaa. Tekstin ymmärtämisen selkeyttämiseksi, englanninkieliset yleisesti käytössä olevat termit esitellään aina, kun suomenkielistä termiä käytetään ensimmäisen kerran. Kyberturvallisuuden termistön suhteen, tässä tutkielmassa on käytetty Turvallisuuskomitean toimeksiannosta rakennettua Kyberturvallisuuden sanastoa (Sanastokeskus TSK, 2018), jonka tavoitteena on standardoida ja vakiinnuttaa alan termistöä Suomen kielelle.

2 Teknologiauhkien välttämisen teoria

Informaatioteknologioita (myöh. IT) voidaan hyödyntää hyvään ja pahaan. Teknologiat itsessään eivät määrittele käyttötarkoitusta, vaan ne toimivat työkaluna, jota käyttäjä voi hyödyntää omien tavoitteidensa saavuttamiseen (Liang & Xue, 2009). Samalla teknologialla voidaan siten saavuttaa suurta hyötyä ja vahinkoa yksilöille, organisaatioille ja jopa maailmantalouden näkökulmasta (Bagchi and Udo 2003). Liangin ja Xuen (2009) teknologiauhkien välttämisen teoria (*Technology Threat Avoidance Theory*, myöh. TTAT) pyrkii selittämään, mitkä tekijät vaikuttavat ihmisen motivaatioon välttää informaatioteknologiaan kohdistettuja teknologisia uhkia. TTAT on monialainen teoria, joka rakentuu muun muassa prospektiteorian (Tversky & Kahneman, 1992), odotusteorian (Vroom, 1964), kybernetiikan teorian (Carver & Scheier, 1982) ja hallintakeinojen teorian (Lazarus 1966; Lazarus & Folkman, 1984) konstruktioiden varaan. TTAT on suunniteltu täydentämään käyttäjien välttämiskäyttäytymiseen liittyvää teoreettista aukkoa, johon on aiemmin pyritty vastamaan epäsoveltuvasti vain teknologioiden hyväksymistä käsittelevien teorioiden avulla.

Tässä sisältöluvussa esitellään teknologiauhkien välttämisen teoria, sen muuttujat ja kehitys. Luku on jaettu yhteensä viiteen osaan. Ensimmäinen alaluku esittelee teorian taustan. Alaluvussa 2.2 esitellään TTAT:hen liittyvät kognitiiviset prosessit, 2.3 käsittelee käyttäjän sosiaalista ympäristöä ja sen vaikutuksia, sekä alaluvussa 2.4 tarkastellaan teorian kehitystä ja lisäyksiä.

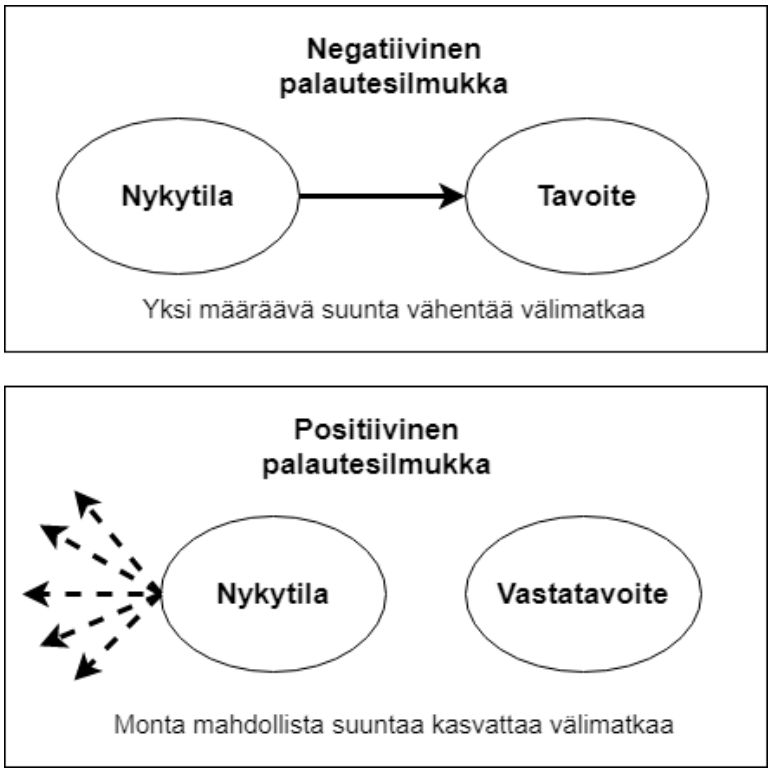
2.1 Tausta

Liang ja Xue (2009) argumentoivat, että ihmisen käyttäytymistä IT-uhkan välttämiseksi ei voida selittää täysin teknologian hyväskymisteorioiden, kuten TAM:n (*Technology Acceptance Model*) ja TRA:n (*Theory of Reasoned Action*), avulla. Liang ja Xue (2009) omaksuvat TTAT:hen kybernetiikan teorian mukaisen näkemyksen, jossa itsereguloivat systeemit, kuten ihmiset, säätelevät käyttäytymistään positiivisten ja negatiivisten palautesilmukoiden avulla (Carver & Scheier 1982). Carver (2006) mukaan hyväksyntä ja välttäminen ovat konseptuaalisesti ja teoreettisesti eri asioita; hyväksyntään liittyy yksilön pyrkimys vähentää nykytilan ja tavoitteen välistä välimatkaa, kun taas välttämisessä ihminen pyrkii kasvattamaan etäisyyttä epämieluisaan vastatavoitteeseen. Hyväksyntään liittyvä kognitiivinen prosessi on siis negatiivinen palautesilmukka, jossa ihminen tavoittelee selkeää päämäärää hyvin rajatuilla toimenpiteillä, ja välttämisessä positiivinen, jossa vastatavoitteen toteutumisen ehkäisemiseksi voidaan toimia hyvin monilla eri tavoilla (Carver, 2006) (ks. kuvio 1).

Liang ja Xue (2009) jakavat teknologiauhkien välttämisen teoriassa IT:t käyttäjien käsityksen perusteella "hyveellisiin" (engl. *virtuous*) ja "vahingollisiin"

(engl. *malicious*). Hyveelliset IT:t ovat käyttäjiensä näkökulmasta puoleensavetäviä sekä aikaansaavat positiivisia lopputulemia. Tähän kategoriaan kuuluvat esimerkiksi perinteiset kehittämiseen ja organisaation liiketoiminnallisiin tarpeisiin rakennetut tietojärjestelmät ja alustat, kuten toiminnanohjausjärjestelmät ja tietokoeavusteiset suunnittelujärjestelmät. Päinvastaisesti, vahingolliset IT:t, kuten vakoiluohjelmat, ovat hyveellisen IT:n käyttäjien näkökulmasta luotaantyöntäviä ja aiheuttavat negatiivisia lopputulemia. Informaatioteknologisessa kontekstissa nämä negatiiviset lopputulemat liittyvät lähinnä tietokoneen suorituskykyyn ja yksityisyyteen. Esimerkiksi tietokonevirus voi olla ohjelmoitu aiheuttamaan toimintahäiriöitä ja varastamaan kohdejärjestelmästä tärkeitä tiedostoja. (Liang & Xue, 2009).

Vahingollisen IT:n tuomien uhkien välttämiseksi ja hyveellisen IT:n prosessien turvaamiseksi voidaan hyödyntää suojaavia toimenpiteitä (engl. *Safeguarding Measures*). Suojaavat toimenpiteet voivat olla IT-järjestelmiä tai käyttäjän tekoja uhkan ehkäisemiseksi. Käyttäjä voi esimerkiksi asentaa tietokoneviruksen tuomaa uhkaa vastaan viruksentorjuntaohjelmiston tai ehkäistä käyttäjätiliensä väärinkäytön mahdollisuutta päivittämällä salasanojaan säännöllisesti. Suojaavien toimenpiteiden avulla käyttäjä pyrkii positiivisen palautesilmukan mukaisesti kasvattamaan etäisyyttään vastatavoitteeseen. (Liang & Xue, 2009).



Kuvio 1 Negatiivisen ja positiivisen palautesilmukan erot (suomennettu Liang & Xue, 2009, s. 76 kuviosta)

2.2 Kognitiiviset prosessit

Käyttäjät käyvät läpi kolme kognitiivista prosessia täydellisessä TTAT:n mukaisessa positiivisessa palautesilmukassa. Prosessit tapahtuvat järjestyksessä ja edellyttävät aina edellisen prosessin suorittamista ennen seuraavan aloittamista (Liang & Xue, 2009). Nämä kognitiiviset prosessit esitellään seuraavaksi samassa järjestyksessä kuin ne käydään läpi käyttäjien toimesta: 2.2.1 Uhkan arviointi; 2.2.2 Hallintakeinojen arviointi; ja 2.2.3 Tilanteen hallinta. Kussakin alaluvussa käydään läpi kyseiseen kognitiiviseen prosessiin liittyvät muuttujat.

2.2.1 Uhkan arviointi

Liang ja Xuen (2009) esittämä TTAT:n mukainen prosessi lähtee liikkeelle, kun käyttäjä tulee tietoiseksi vahingollisesta IT:sta ympäristössään. Hän asettaa vahingollisen IT:n aiheuttaman uhkan realisoitumisen vastatavoitteekseen, joka halutaan välttää. Positiivisessa palautesilmukassa ihmiset käyvät hallintakeinojen teorian (Lazarus 1966; Lazarus & Folkman, 1984) mukaisesti läpi kaksi kognitiivista prosessia päättäessään, aikovatko he toimia IT-uhkan ehkäisemiseksi. Vertaillessaan nykytilannetta ja vastatavoitetta toisiinsa, hän muodostaa subjektiivisen arvionsa vahingollisen IT:n uhkaavuudesta (engl. *Perceived Threat*) kahden muuttujan kautta: mielletty alttius (engl. *Perceived Susceptibility*) ja mielletty vakavuus (engl. *Perceived severity*) (Liang & Xue, 2009).

Liang ja Xue (2009) määrittelevät mielletyn alttiuden käyttäjän arvioimana todennäköisyytenä sille, että uhkan realisoituessa hän kohtaa itse negatiivisia vaikutuksia. Mielletty vakavuus puolestaan mittaa käyttäjän arviota uhkan realisoitumisen aiheuttamien seurausten vakavuudesta. Käyttäjä kokee vahingollisen IT:n sitä uhkaavammaksi, mitä alttiimpi hän uskoo olevansa vahingollisen IT:n negatiivisille vaikutuksille, ja mitä vakavampia hän uskoo seurausten toteutuessaan olevan. Muuttujien välisen vuorovaikutuksen vuoksi, toisen muuttujan saadessa nolla-arvon, koko uhka-arvioyhtälön lopputulos on myös nolla. Eli mikäli käyttäjä esimerkiksi kokee vahingollisen IT:n mahdolliset seuraukset äärimmäisen vakavina, mutta arvioi tapahtuman mahdottomaksi, hän ei miellä vahingollista IT:aa uhkaksi, eikä siten pyri myöskään välttämään sitä. (Liang & Xue, 2009).

Liang ja Xue (2009) omaksuvat TTAT:hen myös uhkan arviointiprosessiin vaikuttavan riskinsietokyvyn (engl. *Risk Tolerance*) muuttujan. He jakavat Barskyn ym. (1997) näkemyksen, jonka mukaan korkean taloudellisen riskinsietokyvyn omaavat ihmiset ovat taipuvaisempia hyväksymään suurempia riskejä myös muissa yhteyksissä. Liang ja Xue (2009) argumentoivat siten, että käyttäjän riskinsietokyky vaikuttaa negatiivisesti hänen arvioonsa vahingollisen IT:n uhkaavuudesta.

2.2.2 Hallintakeinojen arviointi

Hallintakeinojen arviointiprosessin läpikäymisen ehtona on, että käyttäjä on suorittanut uhkan arvioinnin ja kokee vahingollisen IT:n uhkaavaksi (Liang & Xue, 2009) Hallintakeinojen arvioinnissa käyttäjä arvioi uhkan vältettävyyttä (engl. *Perceived Avoidability*) suojaavan toimenpiteen käyttöönoton kautta. Uhkan vältettävyys muodostuu kolmesta eri muuttujasta, jotka ovat mielletty tehokkuus (engl. *Perceived Effectiveness*), mielletyt kustannukset (engl. *Perceived Costs*) ja minäpystyvyys (engl. *Self-efficacy*). (Liang & Xue, 2009)

Liang ja Xue (2009) mukaisesti, käyttäjän mieltämä tehokkuus mittaa, kuinka tehokkaasti hän uskoo suojaavan toimenpiteiden estävän vahingollisen IT:n tuoman uhkan realisoitumisen. Arvioidessaan suojaavan toimenpiteen käyttöönoton kustannuksia, käyttäjä estimoi häneltä vaadittujen fyysisten ja kognitiivisten ponnistelujen määrää, kuten hintaa, ajankäyttöä, epäkäytännöllisyyttä ja ymmärrystä (Weinstein, 1993). Minäpystyvyys puolestaan mittaa käyttäjän itseluottamusta sen suhteen, että hän osaa ottaa suojaavan toimenpiteen asiaankuuluvasti käyttöön. Käyttäjän kokonaisarvio uhkan vältettävyyden asteesta on sitä korkeampi, mitä tehokkaampana hän pitää suojaavaa toimenpidettä uhkaa vastaan, mitä kykenevämpi hän uskoo olevansa sen käyttöönotossa ja mitä alhaisempia siihen liittyvät kustannukset hänen arvionsa mukaan ovat. (Liang & Xue, 2009)

2.2.3 Tilanteen hallinta

Liang ja Xue (2009) omaksuvat Lazarusin ja Folkmanin (1984) hallintakeinojen teorian mukaiset kaksi lähestymistapaa, joiden avulla käyttäjä pyrkii saamaan uhkaavan tilanteen hallintaansa: ongelma- ja tunnelähtöiset hallintakeinot (engl. *Problem- and Emotion-focused Coping*). Uhkan ja hallintakeinojen arviointiprosessien lopputulokset, eli käyttäjän muodostamat käsitykset vahingollisen IT:n uhkaavuudesta ja uhkan vältettävyydestä, toimivat syötteenä tilanteen hallinnan prosessille. Ne määrittävät käyttäjän motivaation (engl. *Avoidance Motivation*) pyrkiä välttämään uhkan negatiiviset vaikutukset, ja siten vaikuttavat hänen hyödyntämiensä hallintakeinojen valintaan.

Mikäli käyttäjä kokee olevansa tilanteessa kontrollissa, hän pyrkii toimimaan ongelmalähtöisesti (engl. *Problem-focused Coping*). Käyttäjä on tällöin motivoitunut välttämään uhkan ja konkreettisesti tavoittelee sitä välttämiskäyttäytymisen (engl. *Avoidance Behavior*) kautta, eli ottaa suojaavia toimenpiteitä käyttöönsä. Ratkaistessaan tilannetta onglemalähtöisesti, käyttäjä pyrkii objektiivisesti vähentämään riskiä altistua vahingollisen IT:n negatiivisille seurauksille kasvattamalla välimatkaa suhteessa vastatavoitteeseen.

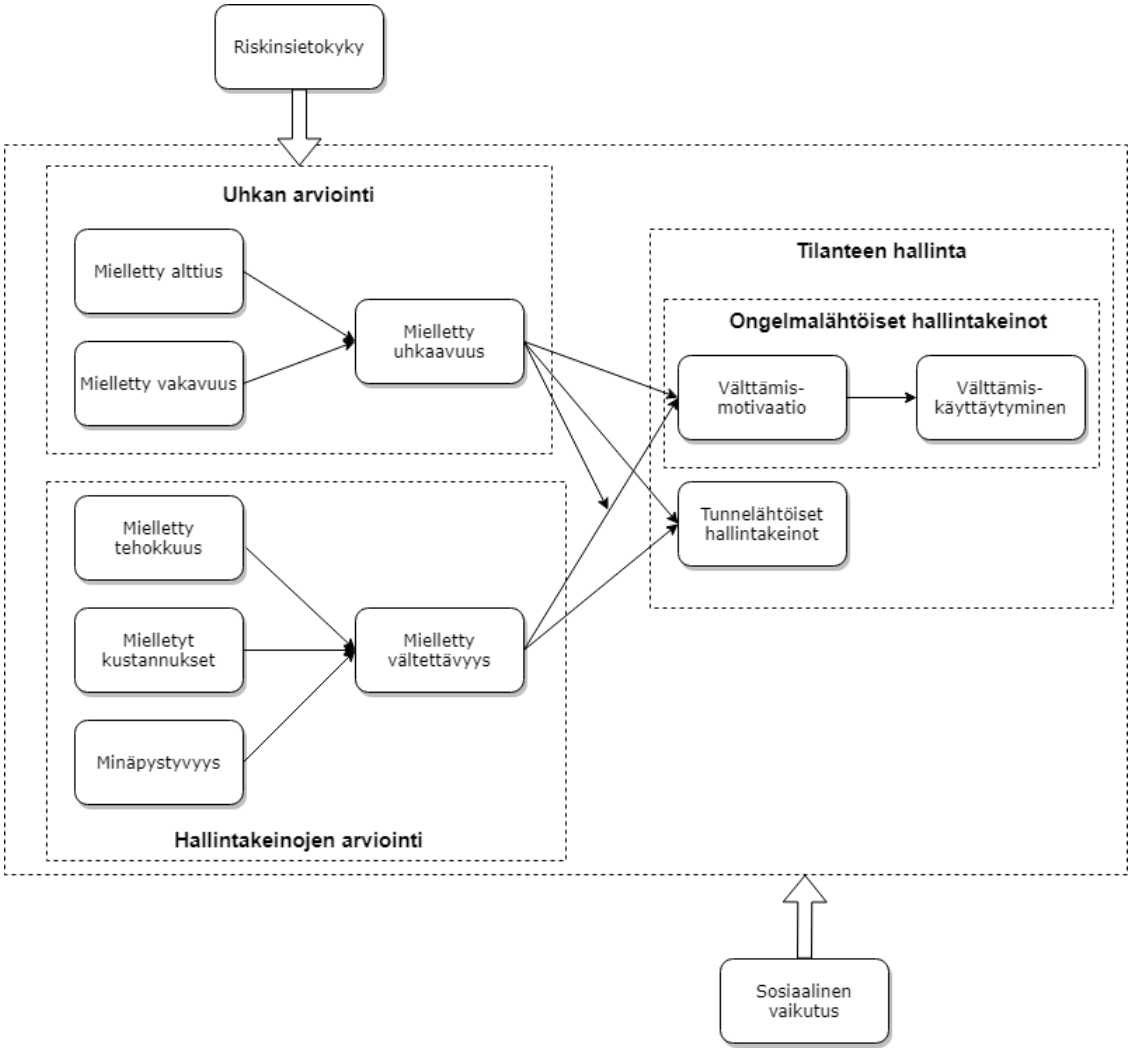
Siinä suhteessa, missä käyttäjä kokee, ettei uhka ole vältettävissä suojaavan toimenpiteen keinoin, hän täydentää hallinnan puutetta tunnelähtöisten keinojen (engl. *Emotion-focused Coping*) avulla. Näitä ovat muun muassa uhkan olemassaolon kieltäminen, uhkan läsnäolon hyväksyminen ja avuttomuuteen heittäytyminen. Tunnelähtöisiä keinoja hyödyntäessään, käyttäjä ei objektiivisesti edistä

vahingollisen IT:n tuoman uhkan poistamiseksi ympäristöstä. Sen sijaan hän muodostaa itselleen vääristyneen kuvan ympäröivästä todellisuudesta, jonka avulla hän subjektiivisesti vähentää uhkan aiheuttamaa pelkoa ja stressiä itsessään, mutta samalla heikentää motivaatiotaan suorittaa ongelmalähtöistä hallintaa. Syinä tunnelähtöisiin keinoihin turvautumisessa voi myös olla käyttäjän havainnot suojaavien toimenpiteiden tehottomuudesta tai niiden täysi puute. (Liang & Xue, 2009)

2.3 Sosiaalinen ympäristö

Käyttäjät toimivat lähes aina osana sosiaalista ympäristöä, kuten ryhmää, organisaatiota ja yhteisöä. Tällöin he ovat väistämättä myös ympäristönsä vaikutuksen alaisia (Deutsch & Gerard, 1955). Käyttäjien on mahdollista täydentää omaa tiedon ja osaamisen puutettaan vastaanottamalla sosiaalisesta ympäristöstään informaatiota, jota he hyödyntävät arvioidessaan vahingollisen IT:n uhkaavuutta ja uhkan vältettävyyttä (Liang & Xue, 2009). Sosiaalisella ympäristöllä on myös normatiivinen vaikutus käyttäjään. Jokaisella käyttäjällä on subjektiivinen näkemyksensä suhteestaan sosiaalisen ympäristöönsä ja siinä toimiviin ihmisiin (Kelman, 1974). Lisäksi käyttäjä voi pyrkiä muokkaamaan oman arvomaailmansa sosiaalisen ympäristönsä arvomaailman mukaiseksi sekä saavuttamaan hyväksyntää ja välttämään normien vastaisen toiminnan aiheuttamia rangaistuksia. Normatiivinen paine johtaa siihen, että käyttäjä saattaa toimia tietyllä tavalla siksi, koska hänen sosiaalinen ympäristönsä vaatii sen. Käyttäjä voi esimerkiksi toimia uhkan välttämiseksi, koska hänen työpaikkansa vaatii sitä, vaikkei hän kokisikaan vallitsevaa tilannetta uhkaavaksi (Kelman, 1974).

Edellä olevaan perustuen, Liang ja Xue (2009) implementoivat TTAT:hen sosiaalisen vaikutuksen (engl. *Social Influence*) muuttujan. He argumentoivat, että käyttäjän sosiaalisella ympäristö vaikuttaa käyttäjän läpikäymien uhkan- ja hallintakeinojen arviointiprosesseihin ja normatiivisilla paineilla on puolestaan suora vaikutus hänen motivaatioonsa toimia uhkan välttämiseksi. Kuvio 2 esittää Liangin ja Xuen (2009) teknologiauhkien hyväksymisen teorian muuttujia ja niiden välisiä tilastollisesti merkitseviä yhteyksiä. Teorian mukaiset kokonaisuudet on rajattu katkoviivalla. Sosiaalinen vaikutus koskettaa koko TTAT-prosessia.



Kuvio 2 Teknologiauhkien välttämisen teorian (TTAT) varianssiteoreettinen näkemys (suomennettu ja muokattu Liang & Xue, 2009, s. 79)

2.4 Teknologiauhkien välttämisen teorian kehitys

Liang ja Xue (2010) pyrkivät validoimaan TTAT:n mukaisen yksinkertaistetun mallin. Tutkimuksen tulokset antoivat teorialle vahvaa tukea muiden paitsi mielletyn alttiuden ja vakavuuden välille ehdotetun yhteisvaikutuksen osalta. Myöhemmät tutkimukset eivät ole myöskään pystyneet osoittamaan näiden teoreettisten konseptien välistä yhteyttä (mm. Arachchilage & Love, 2013; Young, Carpenter, McLeod, 2016) Boysen ym. (2019) ehdottavat tämän ristiriitaisuuden seurauksena muutosta TTAT:n uhkanhallinnan prosessiin. Heidän sekä Carpenterin, Youngin, Barrettin ja McLeodin (2019) tutkimustulokset tukevat ehdotusta, jonka mukaan mielletty vakavuus toimii medioivana muuttujana mielletyn alttiuden ja uhkaavuuden välillä.

Carpenter ym. (2019) tutkimustulokset osoittavat myös tukea kahdelle lisämuuttujalle: riskikäyttäytymis- ja epäluottamustaipumus (engl. *Risk Propensity*,

Distrust Propensity). Sitkiniltä ja Weingartilta (1995) omaksuttu riskikäyttäytymistaipumuksen muuttuja jalostaa samaa ajatusta Liangin ja Xuen (2009) käyttäjän riskinsietokyvystä. Riskikäyttäytymistaipumus tarkoittaa TTAT:n yhteydessä käyttäjän kumulatiivista taipumusta hakeutua riskitilanteisiin tai välttää niitä (Carpenter ym., 2019). Epäluottamustaipumus puolestaan viittaa suojaavaan toimenpiteeseen kohdistuvan luottamuksen puutteeseen sekä odotukseen, että sen kehittäjät, ylläpitäjät tai toimeenpanijat ovat epäpäteviä ja heidän toimintansa turvatonta ja riittämätöntä uhkan välttämiseksi (McKnight & Chervany, 2001; Carpenter ym., 2019).

Muista TTAT-tutkimuksista poiketen Carpenter ym. (2019) eivät löytäneet tilastollisesti merkittävää tukea käyttäjän minäpystyvyyden ja välttämismotivaation välillä. Carpenter ym. (2019) nostaa tämän eroavaisuuden mahdolliseksi syyksi minäpystyvyyden mittarin epäsoveltuvuuden kyseisessä tutkimuksessa.

3 Koneoppiminen

Samuelin (1959) määritelmän mukaan, koneoppimisessa tietokone oppii tehtävän itsenäisesti kokemuksen pohjalta, jolloin järjestelmän toimintaa ei tarvitse erikseen yksityiskohtaisesti ohjelmoida. Tämän prosessin tuotteena syntyy tilastolliseen dataan pohjautuva matemaattinen malli, jonka avulla pyritään yleistämään vastuksia tulevaisuudessa kerättäville havainnoille (Alpaydin, 2010, s.39). Alpaydin (2016) kuvaa koneoppimista prosessina, jossa datasta tuotetaan algoritmien avulla uutta merkityksellistä informaatiota. Koneoppimisen taustalla ovat pyrkimykset mallintaa ihmismäistä oppimisprosessia koneellisesti ja siten automatisoida informaation päättelemisen raakadatasta (Liu, 2018). Jotta järjestelmän voidaan sanoa olevan oppiva, sen suorituskyvyn kehityksen tulee olla kasvujohteista mallin opettamisen edetessä (Michalski, Carbonell, & Mitchell, 1985).

Koneoppimisen sovelluskohteisiin kuuluvat muun muassa hahmon- ja puheentunnistus, luonnollisten kielten prosessointi ja suosittelujärjestelmät (Hastie, Tibshirani & Friedman, 2011). Lisäksi koneoppimista hyödynnetään laajasti myös kyberturvallisuuden edistämisessä, kuten hyökkäyksen havaitsemiseen perustuvissa IDS- ja IPS-järjestelmissä (Buczak & Guven, 2015) sekä loki- ja tapahtumatietoja tallentavissa SIEM-järjestelmissä (Feng, Wu & Liu, 2017). Koneoppimisen viimeaikaisen suosion ja alan kehityksen ovat mahdollistaneet datamassojen saatavuus mallien opettamiseen sekä koneoppimisen kehitykseen kohdistettujen ohjelmointikirjastojen saatavuus. Ennen koneoppimisen käytännöllistymistä monien järjestelmien toiminta oli riippuvaista yksityiskohtaisesta ja käsin tehtävästä ohjelmoinnista. (Alpaydin, 2010, s. 15-16).

Tämä luku tarkastelee koneoppimista teknologiana ja antaa yksinkertaistetun yleiskuvan koneoppimista ympäröivästä pelikentästä. Alaluku 3.1 3.2 koneoppimisen eri menetelmiä; 3.3 kuvailee koneoppimista prosessina; ja 3.4 koneoppimisen tarkkuuteen vaikuttavia tekijöitä. Viimeinen alaluku 3.5 esittelee konkreettisia tapoja kirjallisuudesta, miten koneoppimismenetemiä ja niiden heikkouksia voidaan hyödyntää hyökkäyksissä.

3.1 Oppiminen prosessina

Koneoppimisprosessi voidaan jaotella kahteen päävaiheeseen: opetus- (engl. *training phase*) ja päättelyvaiheeseen (engl. *inference phase*). Opetusvaiheeseen kuuluu datan keräys, esiprosessointi sekä mallin opettaminen oppimisalgoritmin avulla, ja päättelyvaihe käsittää ajanjakson, jona opetettu ja käyttöönotettu malli tuottaa vastauksia uusille havainnoille (Papernot ym., 2016c).

Kaikki koneoppiminen vaatii dataa tutkittavasta ilmiöstä. Yleistämiskyvyn sekä edustavan otannan takaamiseksi, algoritmit vaativat yleisesti ottaen suuren

määrän dataa (Abadi, ym., 2016a). Kerätty ja koottu data muodostaa havaintoaineiston, jonka havaintoja pyritään ryhmittelemään ja erottelemaan piirteillä (engl. *feature*) (Kotsiantis, 2007). Piirteiden arvot toimivat koordinaatteina, joiden avulla havainnon sijainti piirreavaruudessa (engl. *feature space*) voidaan osoittaa. Piirreavaruuden ulotteisuus eli kompleksisuus määräytyy piirteiden lukumäärän mukaan (Kotsiantis, 2007). Havaintoaineiston esiprosessointivaiheessa data valmistellaan opetusvaihetta varten esimerkiksi huolehtimalla, että jokaisen havainnon jokaisella piirteellä on sille soveltuva arvo sekä karsimalla yleistämistarkkuutta heikentävät piirteet datasta pois (Witten ym., 2011, s. 270, 287).

Datan keräyksen ja esiprosessoinnin jälkeen malli opetetaan. Koneoppimisessa keskeinen tavoite on aikaansaada malli, joka kykenee yleistämään vastuksia opetusaineiston ulkopuolisille havainnoille (Alpaydin, 2010, s. 39). Opetusiteraatioissa havaintoaineisto jaetaan opetusaineistoksi ja testidataksi. Opetusdata syötetään oppimisalgoritmille, joka säätää mallin parametrien arvoja. Mallin tarkkuutta ja sen kykyä yleistää arvioidaan testivirheellä, joka kuvastaa mallin virheellisten vastauksien suhteellista osuutta (Papernot ym., 2016b). Koneoppimismallin kyky erotella havaintoja toisistaan systemaattisesti oikein määrittelee siten koneoppimismallin hyödyllisyyden. Koneoppimisprosessin ja malleja kehittävien käyttäjien voidaan siis katsoa toimivan suorituskykykriittisessä ympäristössä.

Mallin opettaminen voi tapahtua offline- tai online-ympäristössä. Offline-opettaminen tapahtuu ennalta kerätyllä datalla. Online-opetuksessa taas mallin parametreja päivitetään uusia havaintoja käsiteltäessä, joten opetus- ja päättelyvaiheet ovat limittyneet. Online-opetus sopii hyvin dynaamisiin ympäristöihin, joissa data muuttuu nopeasti esimerkiksi trendien vaikutuksesta ja siten se mahdollistaa mukautuvien koneoppimismallien kehittämisen. (Barreno ym., 2006).

Päättelyvaiheessa opetus on suoritettu ja malli on otettu käyttöön. Uudet havainnot haetaan datasäiliöstä tai sensoreista, prosessoidaan yhtenevään muotoon ja syötetään mallille. Mallin tuottama tuloste välitetään käyttäjälle tai toiselle järjestelmälle, joka tarvittaessa ryhtyy tarpeellisiin toimenpiteisiin (Papernot ym., 2016b). Esimerkiksi kyberpuolustukseen rakennettu tunkeilijan havaitsemisjärjestelmä (IDS) hälyttää käyttäjää epänormaalin toiminnan yhteydessä tai tiedetyn hyökkäyksen tunnusmerkin havaittuaan (Kim ym., 2018).

3.2 Menetelmät

Koneoppimisen lähestymistavat jaotellaan yleisesti kolmeen tai neljään eri osaan tulkinnasta riippuen: ohjattu oppiminen, ohjaamaton oppiminen, vahvistusoppiminen ja puoliohjattu oppiminen. Näistä viimeisintä ei eritellä lähdekirjallisuudessa aina omaksi lähestymistavakseen (vrt. Portugal, Alencar & Cowan, 2017 ja Papernot ym., 2016b). Lähestymistavan valintaan vaikuttavat merkittävästi muun muassa saatavilla olevan datan määrä, laatu ja hinta, tieto ja ymmärrys käsiteltävästä ongelmasta, sekä mitä hyötyä käyttäjä tavoittelee saavuttavansa koneoppimismallin opettamisella.

Ohjatussa oppimisessa opetusaineiston havainnoilla on olemassa tiedetty vastemuuttujan arvo (Portugal, Alencar & Cowan, 2017). Ohjaamattomassa oppimisessa aineistossa ei ole selitteitä lainkaan. Esimerkiksi, klusterointi on ohjaamattoman oppimisen muoto, jossa tarkoituksena on ryhmitellä keskenään samankaltaiset havainnot, eli piirreavaruudessa lähekkäin olevat havainnot. Vahvistusoppimisessa algoritmille syötettävän opetusaineiston selitteitä ei ole saatavilla, mutta on mahdollista arvioida, oliko muutos mallin suorituskyvyssä toivottu. Oppimisalgoritmi optimoi suorituskykyä kuvaavan virhefunktion arvon, ja tallentaa ne piirre- ja hyperparametriyhdistelmät, joilla saavutettiin maksimaaliset arvot. (Barto & Dietterich, 2004).

Puoliohjattua oppiminen on ohjatun ja ohjaamattoman lähestymistavan yhdistelmä. Sitä sovelletaan usein tilanteessa, jossa opetusaineiston vastemuuttujien arvot tunnetaan vain osalta havainnoista, jolloin dataa ohjattujen menetelmien hyödyntämiseen on liian vähän luotettavien tulosten saavuttamisen kannalta. Puoliohjatun oppimisen taustalla voi olla myös vastemuuttujien arvojen selvittämisen hankkimisen suuri hinta. Päättely puoliohjatuissa menetelmissä perustuu yksinkertaisimmillaan olettamukseen, että piirreavaruudessa lähekkäin sijaitsevat havainnot kuuluvat suurella todennäköisyydellä samaan luokkaan. (Er ym., 2016).

Ympäristöissä, joissa tyypillinen data muuttuu ajan saatossa, kuten IDS-järjestelmissä, voidaan hyödyntää itsenäisesti päivittyviä online-koneoppimismalleja. Mikäli kohdeympäristöstä kerätty data muuttuu ajan saatossa, eri ajanjaksojen havaintojakaumat voivat poiketa niin paljon, että mallin yleistämiskyky kärsii merkittävästi. Esimerkiksi IDS-järjestelmän opetusdatan täytyy vastata ajantasaista verkkoliikennettä, jotta potentiaaliset tunkeutumisyritykset järjestelmään saadaan havaittua ja pysäytettyä. (Barreno ym., 2006; Buczak & Guven, 2015). Online-mallit toisaalta kasvattavat järjestelmän hyökkäysalaa ja usein mahdollistavat muun muassa luokittelumallien toimintaan vaikuttamisen käyttäjän toimesta. Online-mallien ongelmia tarkastellaan lähemmin osiossa 3.4.2.

3.3 Koneoppimismallin yleistämiskyky

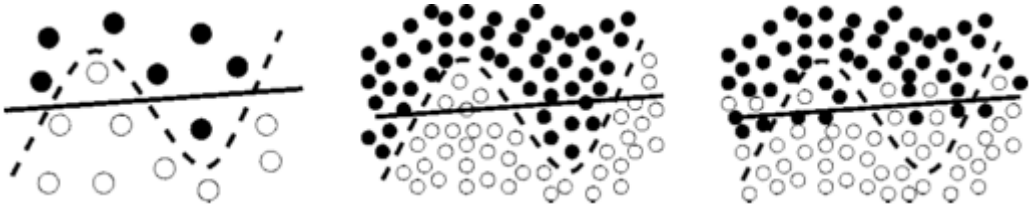
Kaikki koneoppimismallit ovat jollain asteella harhaisia (engl. *bias*) ja olettavat datan havaintojakauman sopivan tietynlaiseen muottiin. Esimerkiksi lineaariset menetelmät olettavat, että datassa olevat havaintoluokat ovat eroteltavissa piirreavaruudessa lineaarisen hypertason avulla ja parametriset menetelmät olettavat tietyn joukon muuttujia, joiden avulla datan rakennetta voidaan todellisuutta kuvaavasti mallintaa (Alpaydin, 2010 s. 38–39). Keskeinen haaste koneoppimisessa onkin löytää ongelman havaintojakaumalle sopiva menetelmä, jonka harha ja olettamukset havaintojakaumasta kuvaavat ympäröivää todellisuutta mahdollisimman hyvin (Alpaydin, 2010 s. 38–39, 60).

Alpaydinin (2010, s. 38–39) mukaan alisovittamisessa (engl. *underfitting*) (ks. Kuvio 3) mallista puuttuu olennaista kompleksisuutta, minkä vuoksi malli ei ”taivu” selittämään havaintoaineiston piilevää monimutkaisempaa rakennetta.

Alisovittumista aiheuttaa muun muassa dataa kuvaavien piirteiden liian vähäinen määrä. Piirteiden määrän kasvaessa mallin kompleksisuus lisääntyy ja havaintoja voidaan erotella yhä hienovaraisempien erojen perusteella (Alpaydin, 2010 s. 38–39). Mikäli mallin kompleksisuus kasvaa suuremmaksi, kuin havaintoaineiston todellinen kompleksisuus, malli ylisovittuu (engl. *overfitting*) (ks. Kuvio 3). Esimerkiksi ylisovittuvassa luokittelumallissa opetusdatalla saavutetaan korkea luokittelutarkkuus, mutta samalla opetusdataan kuulumattoman testidatan luokittelutarkkuus, eli mallin yleistämiskyky, on heikko. (Alpaydin, 2010 s. 38–39).

Havaintojakaumasta poikkeavia havaintoja kutsutaan kohinaksi. Kohinan esiintymisen taustalla voi olla esimerkiksi havaintojen taltiointiin liittyvät epätarkkuudet sekä virheet esiprosessoinnissa tai havainnon selitteissä. Ylisovittuvat mallit oppivat herkästi myös opetusdatan sisältämän kohinan, jolloin päätöspinta ei vastaa ilmiön todellista havaintojakaumaa. Alisovittuvissa malleissa puolestaan puuttuvat piirteet voivat aiheuttaa kohinaksi tulkittavia ilmentymiä havaintoaineistossa (Alpaydin, 2010, s. 30–32; Biggio, Fumera & Roli, 2014).

Suurella joukolla piirteitä on mahdollista oppia monimutkaisempia yhteyksiä havaintojen välillä ja rakentaa huipputarkkoja malleja. Samaan aikaan kuitenkin algoritmien laskennallinen kompleksisuus kasvaa, jolloin mallin opettaminen hidastuu ja vaikeutuu (Zanero & Serazzi, 2008). Yksi koneoppimisen keskeisimmistä haasteista onkin uloitteisuuden kirous (engl. *curse of dimensionality*). Sen mukaan edustavaan otokseen tarvittavan datan määrä kasvaa eksponentiaalisesti suhteessa piirteiden lukumäärään (Bai, 2014). Harhat ja olettamukset johtavat siis epätarkkoihin vastauksiin, kun malli kohtaa niistä poikkeavia havaintoja. Ne ovat kuitenkin olennainen osa mallin ylisovittumisen ehkäisemistä ja yleistämiskyvyn takaamista. Yksinkertaisemmat mallit kykenevät tuottamaan yleisesti ottaen johdonmukaisempia tuloksia komplekseihin ongelmiin ja vaativat vähemmän dataa, kuin monimutkaisemmilla menetelmillä rakennetut mallit, kuten neuroverkot, mutta niiden yleistämiskyky on merkittävästi rajallisempi. (Alpaydin, 2010 s. 32, 38–39).



Kuvio 3 Ali- ja ylisovittaminen kaksiulotteisessa piirreavaruudessa (Müller, Mika, Rätsch, Tsuda & Schölkopf, 2001, 182 kuviosta)

3.4 Vihamielinen koneoppiminen

Vihamieliseksi koneoppimiseksi (engl. A*dversarial Machine Learning*) kutsutaan asetelmaa, jossa koneoppimisprosessin vaikutuspiirissä toimiva taho (myöh.

hyökkääjä) pyrkii aikaansaamaan negatiivisia vaikutuksia koneoppimismallin toimintaan, sen tuottamaan informaatioon tai varastamaan tietoa mallin sisäisestä toiminnasta. Hyökkääjä voi esimerkiksi vääristää mallin vastemuuttujien arvoja tai kerätä tietoa mallin opetusdatasta ja hyödyntää sitä tulevaisuudessa tehtävissä hyökkäyksessä (Barreno ym., 2010).

Hyökkäyssyötteet (engl. *Adversarial Example*) ovat hyökkääjän luomia – synteettisiä tai aidoista havainnoista muuntelemalla saatuja – syötteitä, joiden tavoitteena on luoda uusia tai löytää olemassa olevia syötteitä, jotka malli tulkitsee väärin (Szegedy ym., 2014). Hyökkäyssyötteet voivat sijoittua piirreavaruudessa havaintojakauman ulkopuolelle, jolloin syöte tulkitaan kohinaksi, tai havaintojakauman sisälle, jolloin niiden erot aitoihin syötteisiin voivat olla hyvin pieniä. Havaintojakauman sisälle sijoittuvat hyökkäysyötteet voivat olla erityisen tehokkaita sen vuoksi, että käyttäjän voi olla käytännössä mahdotonta erottaa hyökkäyssyötteitä aidoista syötteistä (Jagielski ym., 2018). Szegedy ym. (2014) osoittavat, että esimerkiksi neuroverkkojen kaltaisissa komplekseissa malleissa hyvin pienet erot piirrearvoissa voivat aiheuttaa riittävän suuria muutoksia ulostulokerroksen vastemuuttujassa, että mallin yleistyskyky heikkenee.

Tässä alaluvussa syvennytään vihamieliseen koneoppimiseen menetelmiin ja niiltä puolustautumiseen. Alaluku sisältää osiot koneoppimisen kohdistettujen hyökkäysten taksonomiasta sekä koneoppimisen hyökkäysalan ja haavoittuvuuksien tarkastelusta koneoppimisprosessin opetus- ja päättelyvaiheissa

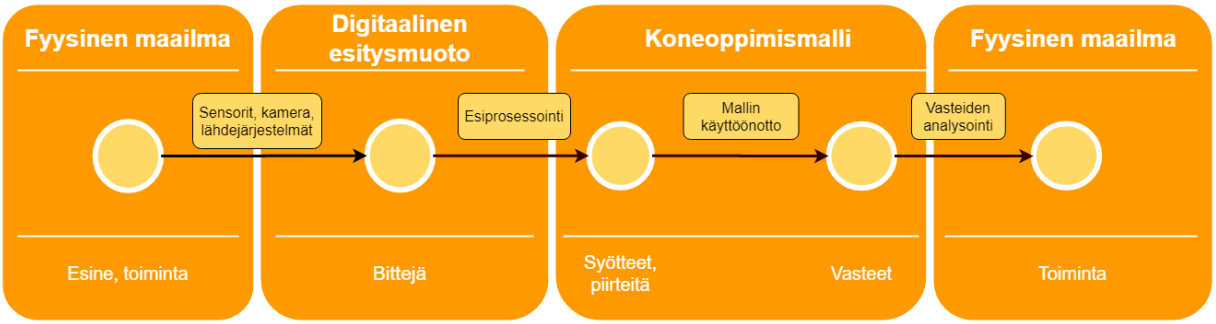
3.4.1 Koneoppimisen hyökkäysala

Koneoppiminen muiden teknologioiden tapaan on altis hyökkäyksille ja väärinkäytölle. Kuten kohdassa 3.1 tarkasteltiin, koneoppiminen on prosessi, johon kuuluvat datan keräys ja prosessointi, mallin opettaminen prosessoidulla datalla sekä tulosteiden analysointi. Nämä osa-alueet muodostavat yhdessä myös koneoppimisjärjestelmän hyökkäysalan (ks. kuvio 4), joka voidaan jakaa prosessin mukaisesti opetus- ja päättelyvaiheiden hyökkäysaloihin.

Papernot ym. (2016) tunnistavat kolme päätyyppiä opetusvaiheen hyökkäyksille: koneoppimismallin kopioiminen, epäsuora vaikuttaminen mallin toiminnallisuuteen ja mallin toimintalogiikan korruptoiminen. Näistä heikoin ja hyökkääjältä vähäisimmän kyvykkyyden vaativa hyökkäystyyppi on mallin kopioiminen. Se vaatii hyökkääjältä pääsyn ainoastaan yksittäisiin mallin opetusdatan havaintoihin tai sen tilastolliseen koosteeseen. Mikäli kerätty data on riittävän laadukasta, hyökkääjä voi onnistua opettamaan varastamallaan opetusdatalla korvikemallin (engl. *surrogate model*). Hyökkääjä voi hyödyntää korvikemallia esimerkiksi hyökkäyssyötteidensä testaamiseen ennen kohdejärjestelmään hyökkäämistä. (Papernot ym., 2016b). Epäsuora vaikuttaminen koneoppimismallin toiminnallisuuteen tapahtuu opetusdatan kautta. Opetusdatan myrkyttämisessä (engl. *poisoning attack*) hyökkääjä saastuttaa opetusdatan tekemillään hyökkäyssyötteitä tai muokkaamalla olemassa olevia havaintoja rikkoen opetusdatan eheyden. Vaikka korvikemallissa käytettävät menetelmät eivät olisikaan

samoja kuin alkuperäisessä mallissa, hyökkäyssyötteiden on havaittu olevan hyvin tehokkaita (mm. Goodfellow, Shlens & Szegedy, 2015; Papernot, McDaniel & Goodfellow, 2016a). Koneoppimisjärjestelmän hyökkäysala kasvaa hyvin usein, kun oppimisprosessi siirretään offline-ympäristöstä online-ympäristöön. Opetusvaiheen hyökkäyksistä loogiseen korruptioon kykenevää hyökkääjää pidetään vahvimpana. Siinä hyökkääjä pystyy muokkaamaan opetusalgoritmin oppimislogiikkaa, kuten hyperparametreja. (Papernot ym., 2016b)

Päättelyvaiheessa hyökkäysten tavoitteena on löytää havaintoja piirreavaruudesta, joissa malli tuottaa virheellisiä tulosteita tai kerätä tietoa mallin sisäisestä toiminnasta. Hyökkäykset päättelyvaiheessa voidaan jakaa white-box ja black-box -menetelmiin, jotka kuvastavat hyökkääjän mahdollisuuksia ja kyvykkyyttä kerätä tietoa koneoppimismallista ja vaikuttaa sen toimintaan (Papernot ym., 2016a). White-box hyökkäyksissä hyökkääjällä on onnistunut omaksumaan tietoa mallin sisäisestä toiminnasta, kuten opetusdatasta, opetusalgoritmista tai sen parametreista. Hyökkääjä pystyy hyödyntämään tätä tietoa esimerkiksi löytääkseen piirreavaruudesta alueita, joissa havainnot luokitellaan herkästi väärin. Black-box -hyökkäyksissä sen sijaan hyökkääjällä ei ole kohteestaan mitään ennakkotietoa, vaan hänen mahdollisuutensa kerätä tietoa mallista ja vaikuttaa sen toimintaan rajoittuu syötteiden ja tulosteiden analysointiin. (Papernot ym., 2016b). Black-box -hyökkäysten ovat havaittu olevan rajoituksistaan huolimatta tehokkaita hyökkäyssyötteiden siirrettävyyden (engl. *adversarial example transferability*) vuoksi. Siirrettävyys tarkoittaa hyökkäyssyötteiden havaittua ominaisuutta toimia eri arkkitehtuuriin perustuvien menetelmien välillä. (Szegedy ym., 2014; Goodfellow ym., 2015).



Kuvio 4 Koneoppimisprosessi ja koneoppimisen hyökkäyspinta (muokattu Papernot ym., 2016b, s. 5 kuviosta)

3.4.2 Hyökkäysten taksonomia

Barreno ym. (2006) ovat luoneet taksonomisen luokittelun koneoppimisalgoritmeihin ja niiden päälle rakennettuihin tietojärjestelmiin kohdistettuja kyberhyökkäyksiä varten (ks. taulukko 1). Hyökkäyksiä voidaan tarkastella kolmesta näkökulmasta: vaikutustapa, tietoturvallisuusloukkauksen tyyppi ja tavoiteltu tarkkuus. Vaikutustavaltaan hyökkäykset jaotellaan kausatiivisiin (engl.

causative) ja tutkiviin (engl. *exploratory*) hyökkäyksiin, tietoturvallisuusloukkauksen pohjalta eheys- ja saatavuushyökkäyksiin sekä tavoitellun laajuutensa mukaan kohdistettuihin (engl. *targeted*) ja kohdistamattomiin (engl. *indiscriminate*) hyökkäyksiin.

Kausatiivisissa hyökkäyksissä hyökkääjällä on mahdollisuus vaikuttaa rakennettavan luokittimen opetusaineistoon. Tavoitteena on johtaa oppimista harhaan niin, ettei malli opi erottelemaan havaintoja tarkoituksenmukaisesti. Tämä voi tapahtua konkreettisesti esimerkiksi vääristämällä opetusaineiston selitteitä tai havaintojakaumaa, jolloin järjestelmä luokittelee hyökkääjän syötteen väärin. Tutkivat hyökkäykset eivät yritä vaikuttaa mallin opetukseen, vaan hyödyntävät sen olemassa olevia puutteellisuuksia. Ne ajoittuvat siis koneoppimisprosessin päättelyvaiheeseen ja niiden tehtävänä on kerätä tietoa algoritmin senhetkisestä tilasta sekä toiminnasta. Tutkivat hyökkäykset pyrkivät paljastamaan mallista haavoittuvuuksia väärin luokiteltujen havaintojen avulla. (Barreno ym., 2006).

Tiedon eheyttä loukkaavat hyökkäykset tavoittelevat tiettyjen koneoppimismallin tulosteiden vääristämistä. Mikäli tavoite on estää laajemmin kohteen pääsy koneoppimismallin parametreihin tai sen tuottamaan merkitykselliseen informaatioon, hyökkäys kohdistuu saatavuuteen (Papernot ym. 2016). Saatavuushyökkäyksen seurauksena koneoppimismallin toiminnasta voi tulla niin epävarmaa, ettei sen varaan rakennetun palvelun tai järjestelmän normaali käyttö ei ole mahdollista. (Barreno ym., 2006; Nelson ym. 2008). Papernot ym. (2016b) määrittelevät luottamuksellisuuden rikkovat hyökkäykset pyrkivät paljastamaan mallin sisäisiä toiminnallisuuksia tai mallin opettamiseen käytetyn aineiston. Mallin sisäinen toiminta voi olla kohteen omistamaa kallisarvoista aineetonta omaisuutta ja siten sen paljastuminen voi aiheuttaa merkittäviä tappioita. Luottamuksellisuuteen kohdistuvalla hyökkäyksellä on myös potentiaali vaarantaa yksityisyys. Esimerkiksi potilastietoa hyödyntävä koneoppimismalli sisältää hyvin sensitiivistä ja levitessään yksityisyyttä loukkaavaa dataa ihmisistä (Papernot ym. 2016b).

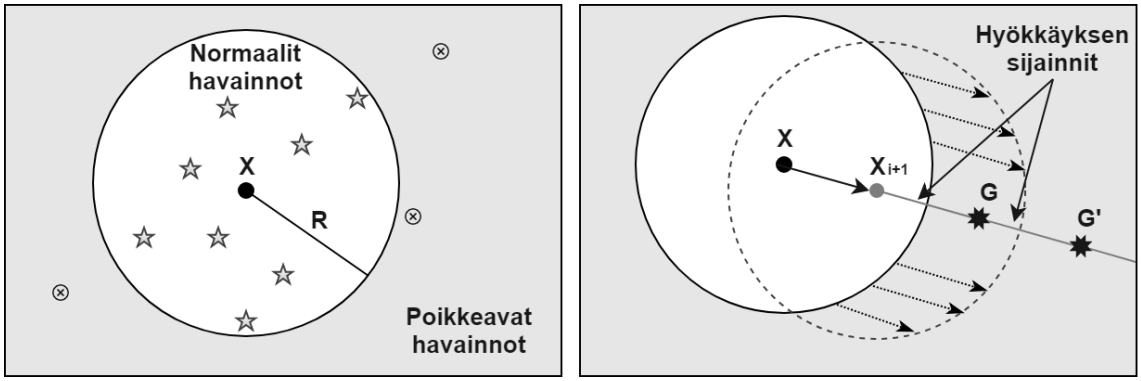
Hyökkäyksen tavoiteltu laajuus tarkoittaa sen havaintojoukkoa, johon hyökkääjä pyrkii vaikuttamaan. Mikäli hyökkääjän aikeena on heikentää luokittelua hyvin tarkasti rajatun havaintojoukon osalta, puhutaan kohdistetusta hyökkäyksestä. Kohdistamattomien hyökkäysten tavoite on puolestaan paljon löyhempi ja hyökkääjälle tärkeämpää on mallin suorituskyvyn yleinen heikentäminen (Barreno, Nelson, Joseph & Tygar, 2010).

		Eheys	***Saatavuus***
Kausatiivinen	*Kohdistettu*	Salli tarkkaan määritetty tunkeutumisen mahdollistava syöte	Saa aikaan riittävästi virheitä, jotta järjestelmästä tulee käyttökelvoton yhdelle ihmiselle tai palvelulle
	Kohdistamaton	Salli ainakin yksi tunkeutumisen mahdollistava syöte	Saa aikaan riittävästi virheitä, jotta koneoppimismallista tulee yleisellä tasolla käyttökelvoton

Tutkiva	*Kohdistettu*	Löydä pienestä potentiaalisten syötteiden joukosta järjestelmän sallima tunkeutumisen mahdollistama syöte	Löydä joukko havaintoja, jotka malli luokittelee väärin.
	Kohdistamaton	Löydä mikä tahansa järjestelmän sallima tunkeutumisen mahdollistava syöte	

Taulukko 1 Koneoppimiseen kohdistettujen hyökkäysten luokittelu (suom. Barreno ym., 2006, s. 3)

Kuten alaluvussa 3.1 esiteltiin, itsenäisesti opetusaineistoa päivittävät online-mallit soveltuvat hyvin dynaamiseen ympäristöön, jossa syötteet muuttuvat ajan kuluessa (Barreno ym., 2006). Kloft ja Laskov (2010) kuitenkin huomauttavat, että tämä lähestymistapa hyvin usein kasvattaa järjestelmän hyökkäyspintaa, ja voi tarjota hyökkääjälle mahdollisuuden vaikuttaa mallin päätöspintaan jopa reaaliaikaisesti. Organisaatioille voi koitua vakavia seuraamuksia, mikäli esimerkiksi sähköpostiohjelman roskapostisuodatin tai IDS-järjestelmä alkaa päästämään hyökkäyssyötteitä läpi. Mikäli dataa ei esiprosessoida ennen uutta opetussykliä, kaikki havainnot päätyvät opetusaineistoon. Tämän vuoksi online-mallit ovat alttiimpia kausatiivisille hyökkäyksille. (Nelson ym., 2008) Kuvio 5 esittää kohdistettua kausatiivista eheyshyökkäystä, jossa hyökkääjä pyrkii siirtämään mallin päätöspintaa ja aikaansaada rakentamiensa havaintojen G ja G' väärinluokittelun. Havaintojakauman keskipisteestä X ulottuva kiinteä säde R määrittää alueen piirreavaruudessa, jossa sen sisällä olevat havainnot (☆) luokitellaan normaaleiksi ja ulkopuolelle jäävät havainnot (⊗) poikkeaviksi. Riittävän kyvykäs hyökkääjä voi siirtää mallin päätöspintaa haluamaansa suuntaan syöttämällä havaintoja systemaattisesti päätöspinnan reunalle ja sen ulkopuolelle.



Kuvio 5 Esimerkki hyökkäyksestä online-luokittimeen (suom. Barreno ym., 2006, s. 7 kuviosta). Vasemmalla lähtötilanne ja oikealla hyökkäyksen aiheuttama päätöspinnan siirtyminen.

3.5 Hyökkäyksiltä puolustautuminen

Koneoppimisen puolustustekniikat pyrkivät säilyttämään mallin ja siihen sidotun informaation tietoturvan. Hyökkäysten negatiivisten vaikutusten minimoimiseksi Barreno ym. (2006) ehdottavat kolmea menetelmää (ks. taulukko 1): regularisointi (engl. *regularization*), satunnaistaminen (engl. *randomization*) ja tiedon piilottaminen (engl. *information hiding*) (ks. taulukko 2). Hyökkääjän yhtenä mahdollisuutena kausatiivisissa hyökkäyksissä on hyödyntää koneoppimismallien kompleksisuutta. Kompleksisuuden vähentämiseksi oppimisprosessiin voidaan implementoida regularisointi, joka rajoittaa mallin kompleksisuutta ja vähentää samalla ylisovittumisen riskiä. Barreno ym. (2006) teoretisoivat kohdistettuja hyökkäyksiä vastaan mahdolliseksi menetelmäksi satunnaistamista. Kohdistetut hyökkäykset tavoittelevat yksittäisten havaintojen väärinluokittelua ja niiden onnistuminen on siten paljon riippuvaisempi päätöspinnan sijainnista piirreavaruudessa, kuin kohdistamattomissa hyökkäyksissä. Mikäli päätöspinnan sijainnissa piirreavaruudessa on satunnaisuutta, hyökkääjä joutuu näkemään enemmän vaivaa onnistuakseen siirtämään päätöspintaa haluamaansa suuntaan. Kääntöpuolena satunnaistamisessa on tarkkuuden heikkeneminen (Barreno ym., 2006). Tutkivia hyökkäyksien tarkoituksena on kerätä tietoa käytössä olevan mallin toiminnasta. Tämän tyyppisiä hyökkäyksiä vastaan Barreno ym. (2006) teoretisoivat koneoppimismallin toimintalogiikkaan, kuten mallin hyperparametreihin ja käytettyihin menetelmiin liittyvän informaation piilottamista. Papernot ym. (2016b) tukevat tätä näkemystä ja osoittavat informaation kasvattavan hyökkääjän kyvykkyyttä rakentaa tehokkaita hyökkäyssyötteitä.

Mallin tarkkuuden ja turvallisuuden välillä on olemassa jännitettä. Mitä herkemmin sovittuva malli on, sitä enemmän informaatiota opetusaineiston havaintojakaumasta opitaan, ja sitä suuremmat mahdollisuudet hyökkääjällä on manipuloida mallia. Mikäli mallin sovittuvuutta rajoitetaan, menetetään myös mahdollisuus oppia tärkeää tietoa uudesta datasta. Mallin joustamattomuus tuo siis turvallisuutta, mutta heikentää sen oppimiskykyä ja käänteisesti joustavuus antaa mahdollisuuden oppia aineistosta enemmän, mutta samalla altistaa mallin hyökkäyksille. (Barreno, 2006).

		Eheys	***Saatavuus***
Kausatiivinen	*Kohdistettu*	• Regularisointi • Satunnaistaminen	• Regularisointi • Satunnaistaminen
	Kohdistamaton	• Regularisointi	• Regularisointi • Satunnaistaminen
Tutkiva	*Kohdistettu*	• Tiedon piilottaminen • Satunnaistaminen	• Tiedon piilottaminen
	Kohdistamaton	• Tiedon piilottaminen	

Taulukko 2 Puolustustekniikoita koneoppimiseen kohdistettuihin hyökkäyksiin (suom. Barreno ym., 2006, s. 4 kuviosta)

Oppimisprosessin onnistumisen ja kausatiivisen hyökkäyksen ehkäisemisen kannalta on oleellista, että hyökkäyssyötteet osataan löytää opetusaineistosta. Löydettyjä hyökkäyssyötteitä voidaan kuitenkin käsitellä eri tavoin. Datan sanitoinnissa (engl. *data sanitization*) eliminoimaan hyökkäyssyötteet aineistosta ja poistaa siten niiden tuoma negatiivinen vaikutus (Liu ym., 2018). Goodfellow ym. (2015) puolestaan tutkivat puolustustekniikkana hyökkäyssyötteillä opettamista (engl. *adversarial training*), jossa hyökkäyssyötteiden pitäminen opetusaineistossa on tarkoituksenmukaista. Taustalla oleva ajatus on, että kun malli oppii erottelemaan hyökkäyssyötteiden piirteet aidoista syötteistä, se pystyy tehokkaasti eliminoimaan hyökkäysten negatiiviset vaikutukset. Goodfellow ym. (2015) toteavat tämän tekniikan olevan tehokas tilanteissa, jossa malli on riittävän kompleksinen oppimaan hyökkäyssyötteiden erityispiirteitä. Mallin kehityksen haasteina kuitenkin ovat sen riippuvuus laadukkaista hyökkäyssyötteistä ja lisääntyneet haasteet ali- ja ylisovittumisen välillä tasapainoilemiseen.

Yhdistelmämetodit (engl. *ensemble methods*) hyödyntävät hyökkäyssyötteiden siirrettävyyttä puolustusmenetelmänä. Tavoitteena on löytää eri malliarkkitehtuurien välillä toimivia hyökkäyssyötteitä, ja sisällyttää nämä mallin opetusdataan (Abbasi & Gagné, 2017; Tramèr ym., 2018). Tramèr ym. (2018) saavuttivat paremman tarkkuuden yhdistelmämetodeilla verrattuna tavallisiin hyökkäyssyötteillä opettamisen metodeihin, mutta osoittavat luokittelutarkkuuden voivan myös kärsiä puhdasta dataa käsiteltäessä. He huomauttavat lisäksi mallien olevan haavoittuvaisia erityyppisiä – mutta silti hyvin yksinkertaisia – hyökkäyksiä vastaan. Tramér ym. (2018)

Papernotin ym. (2016c) tutkimassa ”*defensive distillation*” -tekniikassa syvän neuroverkon resilienssiä hyökkäyssyötteitä vastaan kyetään parantamaan hyödyntämällä mallin itsensä oppimaa tietoa. Malli opetetaan kahdessa iteraatiossa: ensimmäisen iteraation luokkatodennäköisyydet otetaan talteen ja yhdistetään toisessa iteraatiossa opetusaineiston selitteiden kanssa tuomaan luokka-arvoja yksityiskohtaisempaa informaatiota oppimisprosessiin. ”*Defensive distillation*” -tekniikan huomattiin olevan yksinkertainen ja erittäin tehokas piirreavaruudessa aitojen syötteiden lähellä sijaitsevia hyökkäyssyötteitä vastaan, ja onnistui säilyttämään mallin alkuperäisen luokittelutarkkuuden sekä suorituskyvyn (Papernot ym., 2016c).

Yksityisyyden suojaamiseksi koneoppimisprosessi voidaan toteuttaa myös kryptatulla aineistolla. Parhaillaan sekä opetus että päättely voidaan tehdä ilman selkokielisen aineiston käsittelyä säilyttäen mallin alkuperäisen suorituskyvyn. Matemaattisia operaatioita, joita aineistolle voi tehdä on kuitenkin rajatusti, jotta salattu informaatio voidaan vielä palauttaa (Aslett, Esperança & Holmes, 2015; Yao, Song & Chi, 2017).

3.6 Kirjallisuuskatsauksen yhteenveto

Koneoppimisessa pyritään mallintamaan ilmiöitä datan avulla ja tehdä luotettavia päätelmiä ympäristöstä kerätyistä havainnoista. Kaikki koneoppiminen perustuu olettamuksien tekemiseen ongelmasta ja siitä kerättävän datan rakenteesta. Koneoppimisen heikkoudet tulevat esiin, kun data ei vastaa siitä tehtyjä olettamuksia. Hyökkääjät voivat lukuisin eri tavoin estää koneoppimisella tavoitellun hyödyn saavuttamista sekä loukata koneoppimismallin kehitysprosessiin sidotun datan eheyttä, luottamuksellisuutta ja yksityisyyttä, sekä saatavuutta. Vihamielisen koneoppimisen tutkimus on todistanut, että hyökkääjän ei välttämättä tarvitse tietää kohdejärjestelmän sisäisestä toiminnasta mitään suorittaakseen onnistuneita hyökkäyksiä. Lisäksi on osoitettu, että hyökkääjän kyvykkyys koneoppimisprosessin vahingoittamiseen kasvaa merkittävästi hyökkäyksen kohteesta kerätyn informaation kasvaessa. Hyökkäyksille resilienttejä menetelmiä on pyritty aktiivisesti kehittämään vihamielisen koneoppimisen tutkimuksen edetessä ja resilientit menetelmät ovat jo osoittaneet lupaavia tuloksia hyökkäyksiä vastaan. Resilientit menetelmät joutuvat kuitenkin tekemään omat olettamuksensa datan rakenteen lisäksi myös hyökkääjän tavoittelemasta manipulaatiosta ja ovat siten tehokkaita vain, kun olettamukset pitävät paikkansa (Jagielski ym., 2018). Resilienssin rakentaminen koneoppimisprosessiin aiheuttaa siten yhä suuremman haasteen toivotun tarkkuuden saavuttamiselle. Monelle organisaatiolle koneoppimisella on kriittinen rooli liiketoiminnassa. Koneoppimista tulisi suojata sen vuoksi suojata proaktiivisesti.
Teknologiauhkien välttämisen teoria, TTAT (Liang & Xue, 2009), tutkii käyttäjien pyrkimyksiä välttää hyveelliseen informaatioteknologiaan kohdistuvia ja sen tarkoituksenmukaista käyttöä estävien uhkien toteutumista. Käyttäjän päätös hallita uhkaa joko tunnelähtöisesti tai ongelmalähtöisesti riippuu hänen arviostaan uhkan suuruudesta ja käytettävistä keinoista uhkan hallitsemiseen. Töissä käyttäjät toimivat myös osana sosiaalista ympäristöä. Tällöin käyttäjät ovat myös väistämättä niissä vaikuttavien henkilökohtaisten sekä organisaatiossa vallitsevien – toisinaan ristiriitaisten – arvojen, normien ja tavoitteiden vaikutusten piirissä (Deutsch ja Gerard, 1955).

Koneoppiminen tapahtuu TTAT:n (Liang & Xue, 2009) näkökulmasta hyvin mielenkiintoisessa suorituskykykriittisessä ympäristössä. Organisaatiolle olennainen koneoppimismallin arvon määrittävä ominaisuus on sen kyky erotella ja ryhmitellä havaintoja tavoitteellisesti. Mallin suorituskykyä arvioidaan muun muassa testiaineiston testivirheen avulla. Käyttäjän kokeman sosiaalisen paineen ja mahdollisen testivirheen minimoimisen tavoittelun vaikutusta välttämiskäyttäytymiseen ei vielä tunneta.

Koneoppimismenetelmien hyödyntämisellä on organisaatioille erityinen tarkoitus, joka linkittyy organisaatioiden liiketoiminnallisiin tavoitteisiin ja kilpailukyvyn edistämiseen. Koneoppimisen teknologiana voidaan siis tulkita hyveelliseksi IT:ksi. Vihamielisen koneoppimisen menetelmillä (luku 3.4), kuten opetusaineiston myrkyttämisellä ja oraakkelihyökkäyksillä, hyökkääjä pyrkii

vaikuttamaan negatiivisesti kehitettävän mallin tarkkuuteen tai keräämään tietoa sen sisäisestä toiminnasta. Onnistuessaan, yksittäinen hyökkäys voi johtaa merkittäviin rahallisiin menetyksiin, ihmisten yksityisyydensuojan rikkoutumiseen ja jopa kohdeorganisaation liiketoiminnan lakkauttamiseen. Vihamielisen koneoppimisen menetelmien kohdistaminen hyveelliseen IT:aan muodostaa vastatavoitteen ja siten niiden voidaan katsoa lukeutuvan vahingollisten IT:n piiriin. Vihamielistä koneoppimista vastaan rakennettujen puolustusmenetelmien (luku 3.5) voidaan puolestaan katsoa olevan TTAT:ssa suojaavien toimienpiteiden asemassa.

4 Empiirinen tutkimus

Tässä luvussa esitellään tutkielman empiirinen osuus. Ensimmäinen alaluku kertaa tutkimusongelman ja esittelee hypoteesit, toisessa alaluvussa käydään läpi kyselytutkimuksen sisältö, jonka jälkeen viimeisessä alaluvussa esitellään tutkimuksessa käytetty TTAT-malli.

4.1 Tutkimusongelma ja hypoteesit

Tutkielman tavoitteena on tuottaa vastaus tutkielmassa asetettuun tutkimuskysymykseen: *"Mitkä tekijät vaikuttavat välttämiskäyttäytymiseen suorituskykykriittisessä työympäristössä?"*. Tutkielmassa pyritään löytämään mahdollisia eroavaisuuksia välttämiskäyttäytymistaipumuksissa, kun käyttäjä toimii vapaa-ajan sijaan työympäristössä, jossa hänen valintansa vaikuttavat myös organisaation turvallisuuteen. Suorituskykykriittinen ympäristö on tässä tutkielmassa määritelty ympäristönä, jossa työntekijän työpanoksen arvo määrittyy pitkälti muun muassa tuotoksen systemaattisen laadun, tarkkuuden, tehokkuuden tai nopeuden mukaan. Koneoppimista käsittelevässä kolmannessa luvussa, koneoppimismallien kehitysprosessin todettiin tapahtuvan ympäristössä, jossa koneoppimismallin määrittelevänä ominaisuutena on sen kyky erotella havaintoja toisistaan systemaattisesti oikein.

Tutkielma nojautuu vahvasti Liangin ja Xuen (2009) teknologiauhkien välttämisen teoriaan, jota laatiessa on otettu tutkimusongelman kannalta olennainen seikka, vastatavoitteen välttämiseen ja tavoitteen saavuttamiseen liittyvien kognitiivisten prosessien eroavaisuus. Käsitteiden välillä olevat hypoteesit pohjautuvat Boysenin ym. (2019) tutkimukseen, jonka perusteluja hypoteeseille myös tässä tutkimuksessa pitkälti hyödynnetään.

Boysenia ym. (2019) mukaillen tutkimuksen hypoteesit ovat:

H1: Mielletyllä alttiudella (MAL) on positiivinen vaikutus miellettyyn vakavuuteen (MVA)

H2: Mielletyllä vakavuudella (MVA) on positiivinen vaikutus miellettyyn uhkaavuuteen (MUH)

H3: Mielletyllä uhkaavuudella (MUH) on positiivinen vaikutus välttämismotivaatioon (VMO)

H4: Mielletyillä kustannuksilla (MKU) on negatiiviset vaikutukset välttämismotivaatioon (VMO)

H5: Mielletyllä tehokkuudella (MTE) on positiivinen vaikutus välttämismotivaatioon (VMO)

H6: Minäpystyvyydellä (MPY) on positiivinen vaikutus välttämismotivaatioon (VMO)

H7: Sosiaalisella vaikutuksella (SOS) on positiivinen vaikutus välttämismotivaatioon (VMO)

H8: Välttämismotivaatiolla (VMO) on positiivinen vaikutus välttämiskäyttäytymiseen (VKÄ)

4.2 Kyselytutkimus

Kysely toteutettiin Webropol-kyselytutkimusalustalla strukturoituna ja standardoituna 30 kysymyksen lomakkeena, jossa kysymykset ja niiden järjestys olivat kaikille vastaajille samat. Osallistumisen edellytyksenä vastaajilta vaadittiin työkokemusta koneoppimismallien kehittämisestä. Kysely lähetettiin avoimena linkkinä Jyväskylän yliopiston sisäisille opiskelijoiden sekä henkilökunnan sähköpostilistoille. Kohderyhmän tarkan sekä vaativan rajauksen vuoksi, julkinen linkki jaettiin lisäksi LinkedIn-palvelussa, jotta potentiaaliset vastaajat tavoitettiin paremmin. Koska suomenkieliset termit koneoppimisen aihepiiristä eivät ole yhtä tunnettuja ja vakiintuneita kuin englannin kielellä, kysely tehtiin englanniksi. Kielivalinnalla pyrittiin myös mahdollistamaan kansainvälisen vastaajajoukon tavoittaminen.

Kaikki vastaukset kyselyn 28:aan varsinaiseen kysymykseen (ks Taulukko 3) annettiin seitsenportaisella Likert-asteikolla Carpenter ym. (2019) ja Boysen ym. (2019) hyödyntämään tapaan. Likert-asteikon vastausvaihtoehtoina olivat: 1 = vahvasti eri mieltä (*strongly disagree*), 2 = eri mieltä (*disagree*), 3 = jokseenkin eri mieltä (*somewhat disagree*), 4 = en osaa sanoa (*undecided*), 5 = jokseenkin samaa mieltä (*somewhat agree*), 6 = samaa mieltä (*agree*), 7 = vahvasti samaa mieltä (*strongly agree*). Seitsenportainen asteikon valinnalla pyrittiin saavuttamaan tulosten vertailukelpoisuuden lisäksi myös minimaalinen määrä asteikon keskivaiheen ”en osaa sanoa” -vastauksia.

Kyselyn 28:n väittämän kielellisen selkeyden ja ymmärrettävyyden edesauttamiseksi kyselyn ensimmäisellä sivulla määriteltiin kaksi apukäsitettä: hyökkäys (engl. *attack*) ja puolustustekniikat (engl. *defensive techniques*). Näiden käsitteiden merkitys kyselyssä, sekä annetut esimerkit mahdollisten terminologisten väärinymmärrysten ehkäisemiseksi, luotiin tämän tutkielman lukuihin 3.4 ”Vihamielinen koneoppiminen” ja 3.5 ”Hyökkäyksiltä puolustautuminen” kootun materiaalin mukaisesti.

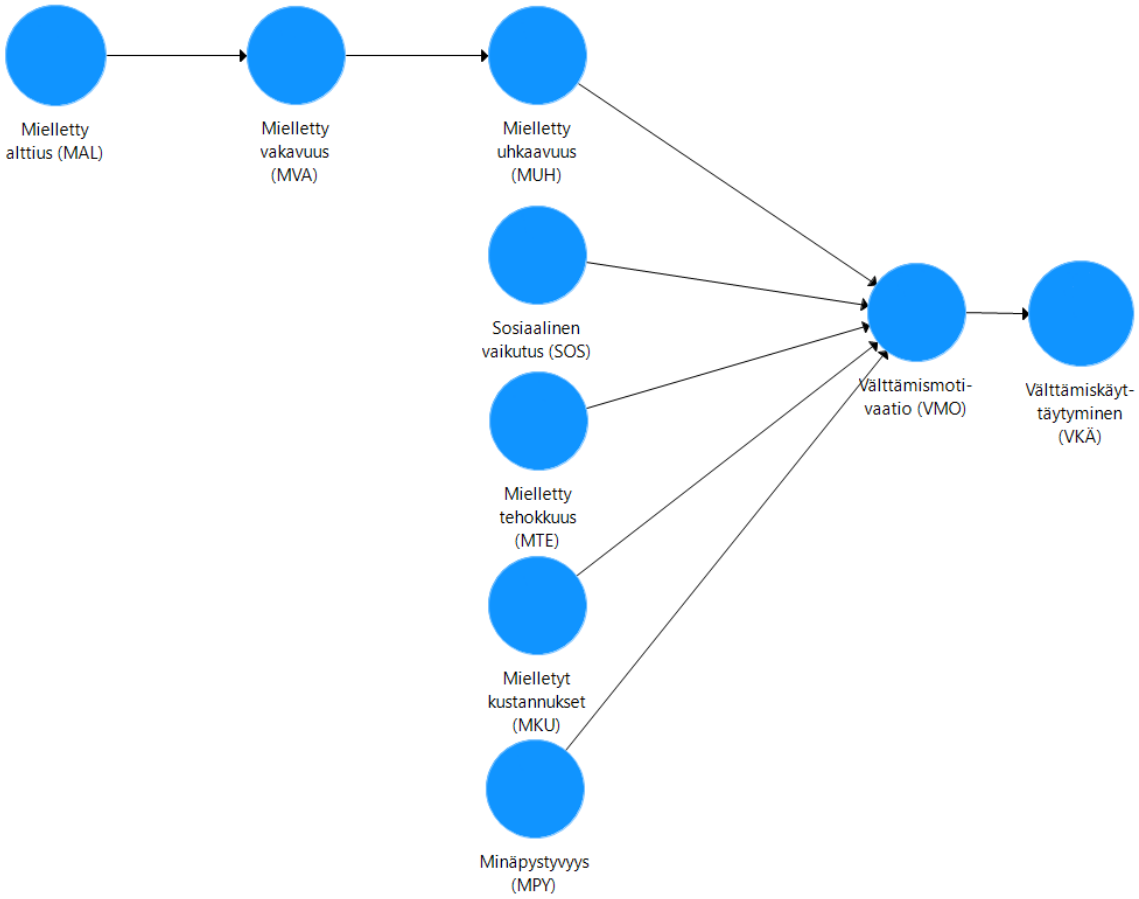
Taulukko 3 Kyselytutkimus

Latentti muuttuja	Koodi	Kysymys
mielletty alttius	MAL1	It is extremely likely that my Machine Learning models, their data, or the outputs produced by them, will be attacked by an adversary in the future
	MAL2	There is a good possibility that my Machine Learning models, their data, or the outputs produced by them will be attacked by an adversary at some point
	MAL3	There is a good chance that my Machine Learning models, their data, or the outputs produced by them, will be attacked by an adversary at some point in the future.
mielletty vakavuus	MVA1	Observations in my Machine Learning model's training data could be added, deleted or perturbed without my knowledge
	MVA2	The information collected from my Machine Learning models could be misused by an adversary.
	MVA3	An adversary's attack could deteriorate the generalization performance of my Machine Learning models, preventing their intended use.
mielletty uhkaavuus	MUH1	An adversary's attack targeted to my Machine Learning model poses a threat to me.
	MUH2	It would be dreadful if the integrity of information in my Machine Learning model was violated by an adversary's attack
	MUH3	It would be risky to use my Machine Learning model if it had been attacked by an adversary.
mielletty tehokkuus	MTE1	Defensive techniques would increase my ability to protect my Machine Learning models from an adversary's attack.
	MTE2	Defensive techniques would enable me to train Machine Learning models, with sufficient generalization performance, faster.
	MTE3	Defensive techniques would increase my productivity in training sufficiently performing Machine Learning models under the presence of an adversary.
mielletyt kustannukset	MKU1	I don't utilize defensive techniques on my Machine Learning models because I am not familiar with them.
	MKU2	I don't utilize defensive techniques on my Machine Learning models, because I'm not familiar with the aspects that need to be considered in their use.
	MKU3	I don't utilize defensive techniques on my Machine Learning models because they complicate the learning process.
	MKU4	I don't utilize defensive techniques on my Machine Learning models because it's more difficult to achieve a high enough performance metric score that is expected from me (e.g. recall, accuracy, F1 score).

sosiaalinen vaikutus	SOS1	My superiors expect that my Machine Learning models should achieve a high score on chosen performance metric (e.g. recall, accuracy, F1 score)
	SOS2	My peers think that I should utilize defensive Machine Learning techniques to protect my models from attacks.
	SOS3	In general, people have supported/recommended the use of defensive Machine Learning techniques.
minäpystyvyys	MPY1	I could successfully implement defensive Machine Learning techniques to my Machine Learning models even if there was no one around to tell me what to do.
	MPY2	I could implement defensive Machine Learning techniques to my models if I had a lot of time to complete the task.
	MPY3	I could implement defensive Machine Learning techniques to my models if someone showed me how to do it first.
välttämismotivaatio	VMO1	I intend to use defensive Machine Learning techniques to avoid adversary's attacks to my models.
	VMO2	I plan to use defensive Machine Learning techniques to avoid adversary's attacks to my models.
	VMO3	I research information about defensive Machine Learning techniques regularly.
	VMO4	I research information about the vulnerabilities of Machine Learning regularly.
välttämiskäyttäytyminen	VKÄ1	I use defensive Machine Learning techniques to identify potential attacks and mitigate their effects on my models.
	VKÄ2	I avoid utilizing Machine Learning techniques that have been found prone to attacks.

4.3 TTAT-malli

Tutkielmassa rakennettu TTAT-malli pohjautuu Boysenin ym. (2019) muunnettua malliin. Mallin latentteja muuttujia kuvaavien indikaattorien määrää jouduttiin karsimaan, jotta vastaajamäärä ei kärsisi sen vuoksi. Käyttäjää itseänsä koskettavia latentteja muuttujia olivat mielletty alttius (MAL), mielletty vakavuus (MVA), mielletty uhkaavuus (MUH), mielletty tehokkuus (MTE), mielletyt kustannukset (MKU) ja minäpystyvyys (MPY). Näiden lisäksi tarkasteltiin myös käyttäjään vaikuttavaa ulkoista muuttujaa sosiaalista vaikutusta (SOS). Kaikki edellä mainitut muuttujat ovat kytketty välttämismotivaatioon (VMO), joka puolestaan kytkeytyy välttämiskäyttäytymiseen (VKÄ). (ks. Kuvio 7).



Kuvio 6 Tutkielman malli Boysenia ym. (2019, s. 102) mukaillen

4.4 Indikaattorien reflektiivinen ja formatiivinen mallinnus

Tämän tutkielman merkittävämpänä erona suhteessa aiempaan TTAT tutkimukseen on latenttien muuttujien ja indikaattorien välisten kausaalisuhteiden suunta. Toisin, kuin mm. Boysenin ym. (2019) sekä Carpenterin ym. (2019) tutkimuksissa, joissa indikaattorit mallinettiin reflektiivisinä, koneoppimisen kontekstissa tämä ei ole mahdollista, niin ettei samalla menetetä oleellista informaatiota. Hulland (1999) kuvaa reflektiiviset indikaattorit kohteina, joihin latentin muuttujan muutokset heijastuvat, kun taas formatiiviset indikaattorit muodostavat yhdessä latentin muuttujan. Howellin, Breivikin ja Wilcoxin (2007) mukaan reflektiivisiä indikaattoreita voidaan arvioida irrallisena ympäröivästä kontekstista niin, että sen nimellinen ja empiirinen merkitys pysyvät linjassa. Formatiiviset indikaattorit puolestaan ovat aina riippuvaisia muista indikaattoreista, eli latentin muuttujan empiirinen merkitys muuttuu sitä ympäröivän kontekstin ja kokonaisuuden muuttuessa (Howell ym., 2007).

Tämän tutkielman kolmannesta luvusta käy ilmi, että koneoppimismallin kehittäminen on hyvin laaja ja moniulotteinen prosessi, jossa lopputuloksena syntyvän mallin hyökkäysala määrittyy usean osa-alueen yhteisvaikutuksen tuloksena. Näihin lukeutuvat käytetyt menetelmät, kuten ohjattu tai ohjaamaton

oppiminen, käytetyt algoritmit, opetustapa, sekä hyökkääjän kykenevyys, kuten hänen omaamansa informaatio kohteesta (ks. Taulukot 1 ja 2). MacKenzie, Podsakoff ja Jarvis (2005) nostavat myös esiin, että reflektiivinen mallinnustapa ja indikaattorien korkeaan yhteneväisyyteen painottunut validointiprosessi johtavat herkästi tilanteeseen, jossa muuttujan konseptuaalinen kenttä katetaan kapeasti ja epätäydellisesti. Tämä on suuri ongelma erityisesti silloin, kun luonteeltaan formatiiviset muuttujat mallinnetaan reflektiivisesti ”väkisin”, tai jos perinteisiä validius- ja reliabiliteettitestejä hyödynnetään formatiivisiin indikaattoreihin (MacKenzie, Podsakoff & Jarvis, 2005). Näihin syihin vedoten argumentoin, että koneoppimiseen liittyviä TTAT-teorian latentteja muuttujia kuuluisi ehdottomasti mallintaa formatiivisesti.

Kyselytutkimus ehdittiin kuitenkin avata vastauksille ennen kuin kirjoittaja tuli tietoiseksi edellä mainitusta koneoppimisen vaatimasta mallinnustavasta. Kun erehdys huomattiin, kyselyyn oli kertynyt riittävän monta vastausta, että mahdollisen kyselyn muokkaamisen ja uudelleenlähettämisen arvioitiin johtavan vain suurempiin ongelmiin, mikäli edes osa kyselyyn vastanneista olisi ollut haluttomia vastaamaan uudistettuun kyselyyn. Sen sijaan kyselyyn päätettiin kerätä loput vastaukset normaalisti ja käsitellä indikaattorit analyysiprosessissa kirjallisuuden ohjeistamalla tavalla.

Kyselyn sulkemisen jälkeen tarkastettiin mallin latenttien muuttujien indikaattorit. Muuttujan MAL (mielletty alttius) indikaattorit mittaavat kaikki käyttäjän kokemaa todennäköisyyttä hyökkäyksen onnistumisesta. Korrelaatiomatriisin korkeat positiiviset arvot antavat myös tukea tälle tulkinnalle konseptuaalisesta päällekkäisyydestä. Samaan tapaan keskenään yhteneviä indikaattoreita havaittiin olevan muuttujissa MKU ja VMO: MKU1 ja MKU2 kuvaavat käyttäjän puolustusmenetelmiä koskevan tiedon puutetta, sekä VMO1 ja VMO2 käyttäjien omista aikomuksista uhkien välttämisen suhteen. Konseptuaalisesti päällekkäiset indikaattorit eivät tuo lisäinformaatiota malliin; päinvastoin korkea indikaattorienvälinen korrelaatio aiheuttaa formatiivisesti mallinnetuissa malleissa herkästi ongelmia epävakaiden regressiopainoarvojen kautta (Cenfetelli & Bassellieri, 2009). Näihin edellä esitettyihin eroihin pohjaten sekä indikaattorien kysymyksenasettelun perusteella todeta, että indikaattorit MAL1, MAL2 ja MAL3, MKU1 ja MKU2 sekä VMO1 ja VMO2 ovat muodostettu reflektiivisesti ja loput indikaattoreista formatiivisesti.

Indikaattorien tarkastuksen yhteydessä huomattiin myös, että sosiaalisen vaikutuksen indikaattorin SOS1 sanallinen muotoilu on päinvastainen suhteessa hypoteesin H7 vaikutuksen suuntaan. Indikaattori kuvastaa käyttäjän kokemaa suoriutumispainetta saavuttaa lunastaa häneen asetetut odotukset. Koska hypotetisoitu vaikutus on positiivinen ja indikaattorin arvon kasvaessa odotettu vaikutus välttämismotivaatioon on negatiivinen, indikaattorin arvot muunnettiin päinvastaisiksi suhteessa Likert-skaalan keskimmäiseen arvoon (7 → 1; 6 → 2; 5 → 3).

5 Analyysi

Tutkielman kuudes luku käsittelee kyselytutkimuksesta kerätyn aineiston analysointia. 5.1 esittelee tilastollisten menetelmien hyödyntämistä edeltävät datan käsittelyt, alaluvussa 5.2 käydään läpi lyhyesti reflektiivisen ja formatiivisen mallintamisen aiheuttamat eroavaisuudet aineiston analysoinnin suhteen.

5.1 Datan käsittely

Aineiston analyysissa hyödynnettiin SmartPLS-ohjelmistoa (SmartPLS, 2020) sekä PLS-regressiota ja bootstrap-menetelmää. Nämä työkalut ja menetelmät mahdollistavat normaalijakaumasta poikkeavan datan analysoimisen (Vinzi, Trinchera, & Amato, 2010). Otoksen kuudentoista vastauksen suuruinen aineisto osoittautui kuitenkin liian pieneksi suhteessa indikaattorien lukumäärään, jotta olisi pystytty hyödyntämään SmartPLS-ohjelmiston bootstrap-menetelmää muun muassa tilastollisten testien merkitsevyyden arvioimiseen. Aineiston kasvattaminen luonnollisesti keräämällä lisää vastaajia ei ollut mahdollista aikapaineen sekä kohderyhmän vaikean tavoitettavuuden yhdistelmän vuoksi, joten tutkimuksessa päädyttiin ratkaisuun kaksinkertaistaa aineiston kaikki kuusitoista kerättyä vastausta. Tälle menetelmälle ei löytynyt tukea kirjallisuudesta, sillä vastaavaa esimerkkiä bootstrap-menetelmän hyödyntämisestä dataan, jossa vastausten määrä on riittämätön, ei löytynyt. Aineisto haluttiin moninkertaistaa täydellisenä, jotta säilytetään havainnoille sama todennäköisyys tulla valikoiduksi bootsrap-otokseen – bootstrap-menetelmä arpoo havaintoja otokseen aina koko aineistosta (Efron & Tibshirani, 1993, s. 13). Havaintoaineiston kaksinkertaistamisen vaikutusta testien tilastolliseen merkitsevyyteen ei kuitenkaan osata arvioida, ja siten tutkimustulosten yleistävyyteen tullaan suhtautumaan varoen. Tutkielman luonteen vuoksi haluttiin joka tapauksessa raportoida p-arvot tulevia tutkimuksia varten ja tarjota siten edes suuntaa antavat arvot otoksen tilastollisesta merkitsevyydestä. Otoksen pienestä koosta ja latenttien muuttujien verrattain pienestä indikaattorimäärästä johtuen, tehtiin myös päätös, että mallia muokataan vain niiltä osin, kuin se on välttämätöntä.

5.2 Formatiivistesti mallinnettujen muuttujien analysointi

Latenttien muuttujien formatiivisuus aiheuttaa sen, että tilastollisia testejä ei voida suorittaa kuten Boysenin ym. (2019) ja Carpenterin ym. (2019) tutkimuksissa, joissa indikaattorien ja latenttien muuttujien välinen suhde on oletettu reflektiiviseksi. Samaa latenttia muuttujaa kuvaavien formatiivisten indikaattorien välillä korrelaatio voi olla positiivisen lisäksi myös negatiivista, mutta sitä ei ole välttämätöntä olla lainkaan. Siinä missä reflektiiviset indikaattorit ovat samasta

latentista muuttujasta riippuvia selitettäviä muuttujia, formatiiviset indikaattorit ovat riippumattomia muuttujia, jotka muodostavat kokonaisuutena latentin muuttujan (Bollen & Lennox, 1991). Ei siis ole mielekästä hylätä formatiivisia indikaattoreita mallista esimerkiksi reliabiliteetin ja konvergentin validiteetin nimissä niiden alittaessa perinteiset 0,70:n Cronbachin alfan sekä faktorilatauksien raja-arvot, sillä formatiivistesti mallinnetun latentin muuttujan ei tarvitse olla sisäisesti yhtenevä, eikä indikaattorien korkeasti korreloivia (Hulland, 1999).

Tutkielmassa muodostetun formatiivisen mallin analysointi suoritetaan soveltuvilta osin Cenfetellin ja Bassellierin (2009) ohjeistusta hyödyntäen. Prosessin tarkoituksena on latenttien muuttujien indikaattorien validiteetti sekä varmistaa mallin teoreettinen ja empiirinen yhtenevyys. Koko prosessia ei valitettavasti pystytä hyödyntämään analyysissa täydellisenä muun muassa indikaattorien rajallisen määrän vuoksi. MacKenzie ym. (2005) huomauttavat, että indikaattorijoukon muutokset voivat vaikuttaa latentin muuttujan konseptuaaliseen kattavuuteen. Indikaattorimäärän rajaus kyselytutkimusta edeltävästi asettaa siten haasteen latenttien muuttujien validiuden saavuttamiselle, mikäli indikaattorit eivät onnistuneesti kuvasta latenttia muuttujaa.

Cenfetellin ja Bassellierin (2009) prosessia mukaillen mallista tarkastetaan ensimmäisenä multikollineaarisuudet. Siinä missä reflektiivisissä malleissa indikaattorien suuri keskinäinen korrelaatio on olennainen ominaisuus, formatiivissa malleissa asia on toisin (Cenfetelli & Bassellierin, 2009). Formatiivisissa malleissa multikollineaarisuus tulee tulkita potentiaalisena konseptuaalisena päällekkäisyytenä, joka ei tuo lisää informaatiota malliin. Vahvasti korreloivista muuttujista voidaan poistaa toinen, mikäli konseptuaalinen päällekkäisyys voidaan osoittaa. Tällä tavoin voidaan pienentää epävakaiden regressiopainotusten riskiä. (Cenfetelli & Bassellieri, 2009; Diamantopoulos & Winklhofer 2001). Multikollineaarisuuden toteamisessa voidaan hyödyntää VIF-arvoja (*Variance Inflation Factor*). Hyväksyttävien VIF-arvojen määritelmistä on olemassa jonkin verran erimielisyyttä tieteellisessä yhteisössä (Cenfetelli & Bassellieri, 2009). Tässä tutkimuksessa hyödynnetään Diamantopoulosin ja Siguawin (2006) ehdottamaa 3,33 raja-arvoa, jonka alle hyväksyttävät arvot sijoittuvat.

Multikollineaarisuuksien tarkastamisen ja mahdollisten muutosten jälkeen analysoidaan indikaattorien absoluuttiset ja suhteelliset vaikutukset latenttiin muuttujaan. Cenfetellin & Bassellierin (2009) mukaan tutkijan tulee raportoida ja tulkita indikaattorien ja latenttien muuttujien väliset absoluuttiset ja suhteelliset vaikutukset, eli indikaattorien faktorilatautumiset sekä PLS-regression tuloksena saatavat osittaiset painoarvot. Latentin muuttujan ja sen indikaattorijoukon teoreettisen soveltuvuuden vahvimpana ilmaisimena toimivat tilastollisesti merkitsevien tulosten ($p < 0,05$) ilmeneminen samanaikaisesti niin absoluuttisissa kuin suhteellisissakin vaikutuksissa. Indikaattorien validiutta kuitenkin mittaavat parhaiten regressiopainoarvot, jotka mittaavat kykyä selittää latentin muuttujan varianssia teoreettisessa kontekstissaan (Petter, Straub & Rai, 2007).

Cenfetelli ja Bassellieri (2009) korostavat, että tilastollisesti merkitseviin tuloksiin yltämättömät vaikutukset eivät kuitenkaan ole suora osoitus indikaatto-

rien tarpeettomuudesta. Tilastollisesti merkitsevän painoarvon yhteydessä ilmenevä ei-merkitsevä faktorilatautuminen voi olla seurausta siitä, että tarkasteltavan indikaattorin vaikutukset latentissa muuttujassa eivät ylitä muiden indikaattorien aikaansaamia vaikutuksia. Formatiiviset indikaattorit käytännössä kilpailevat keskenään latentin muuttujan varianssin selittämisestä, joten vain rajallinen lukumäärä indikaattoreita voi olla tilastollisesti merkitseviä (Cenfetelli & Bassellieri, 2009).

Regressiopainoarvoltaan ei-merkitsevällä indikaattorilla voi myös olla samanaikaisesti merkitsevä faktorilatautuminen, joka on viite indikaattorin kyvystä selittää latentin muuttujan varianssia. Nämä muuttujat ovat olennaista säilyttää mallissa sekä analysoida indikaattorin käyttäytymistä eri konteksteissa (Petter ym., 2007). Mikäli sekä absoluuttisten että suhteelliset vaikutukset eivät ole tilastollisesti merkitseviä, indikaattorin tarpeellisuus mallissa tulisi kyseenalaistaa ja mahdollisesti poistaa heikon tuen vuoksi (Cenfetelli & Bassellieri, 2009). Indikaattorien vähäisen määrän vuoksi, tässä tutkielmassa malliin pyritään tekemään vain ehdottoman tarpeellisia muutoksia. Tilastollisesti ei-merkitsevät tulokset kuitenkin huomioidaan analyysissa ja lopullisesta mallista tehtävissä johtopäätöksissä.

6 Tulokset

Tässä luvussa kuvaillaan kyselytutkimuksen tulokset. Kyselyyn kerääntyi ennen linkin sulkemista yhteensä 16 vastausta, joista kaikkia pystyttiin hyödyntämään analyysissa. Datassa ei ilmennyt puuttuvia havaintoja. Kyselyn ensimmäiset kaksi kysymystä olivat taustakysymykset vastaajan iästä ja kokemuksesta. Vastaajista puolet ovat iältään 30–34-vuotiaita, neljännes 20–24-vuotiaita sekä viimeinen neljännes jakautui tasan 25–29 ja yli 45-vuotiaiden kesken. Vastaajista hieman alle kolmannes vastasi omaavansa hyvin vähäistä työkokemusta koneoppisen käytöstä. Yhdestä kolmeen vuoteen työkokemusta oli 38 prosentilla ja neljästä kuuteen vuoteen neljänneksellä vastaajista. Kahdeksan prosenttia ilmoitti hyödyntäneensä koneoppimista työelämässä vähintään kymmenen vuotta.

Cenfetellin ja Bassellierin (2009) prosessia mukaillen luvussa 6.1 tarkistetaan muuttujien multikollineaarisuudet indikaattorien välillä ja luvussa 6.2 analysoidaan indikaattorien absoluuttiset ja suhteelliset vaikutukset latenttiin muuttujaan. Lisäksi alaluvussa 6.3 käydään läpi polkuanalyysin tulokset sekä niiden implikaatiot mallin hypoteeseihin.

6.1 Multikollineaarisuus

PLS-regression ajamisen jälkeen saatiin yhtenä tuloksena mallin indikaattorien VIF-arvot. Suurta multikollineaarisuutta löytyi odotettavasti reflektiivisesti mallinnetusta muuttujasta MAL. MAL:sta päätettiin luopua kokonaan sen puutteellisen määrittelyn vuoksi, ja koska MAL:n vaikutukset sisältyvät latenttiin muuttujaan MUH (*mielletty uhkaavuus*). Indikaattorien suuri keskinäinen korrelaatio olisi tarkoittanut sitä, että muuttujaa olisi pitänyt kuvata yhden indikaattorin avulla. Vihamielistä koneoppimista käsittelevän luvun 3.4 perusteella voidaan todeta, että koneoppiminen on hyvin moniulotteinen prosessi, jonka eri vaiheita varjostavat erilaiset uhkat. Siten myös käyttäjän näkemys koneoppimisprosessin alttiudesta hyökkäyksille koostuu monesta osasta, eli vain yhden näkökulman sisällyttäminen malliin on riittämätön konseptuaalisella tasolla.

Myös latenteista muuttujista VMO ja MAL todettiin yli 3,33 suuruisia VIF-arvoja, jotka aiheutuivat aiemmin havaitusta konseptuaalisesta päällekkäisyydestä. VMO1 ja VMO2, sekä MKU1 ja MKU2 ovat keskenään hyvin samankaltaisia: indikaattorien konseptuaalisilla merkityksillä on vain marginaaliset erot, jotka myös välittyvät myös korrelatiomatriisin arvoista. Cenfetellin ja Bassellierin (2009) mukaisesti, indikaattorit MKU1 ja VMO2 sekä latentti muuttuja MAL kokonaisuudessaan poistettiin mallista. Muutoksen jälkeen ajettu PLS-algoritmi tuotti kaikille jäljellä oleville indikaattoreille VIF-arvot väliltä 1,003–2,830

eli muuttujien indikaattorien välillä ei vallitse enää multikollineaarisuutta. Alkuperäisen sekä muutosten seurauksena muodostuneen karsitun mallin VIF-arvot ovat nähtävissä taulukossa 4.

Taulukko 4 VIF-arvot alkuperäiselle sekä karsitulle mallille.

	VIF-arvot	
Indikaattori	**Alkuperäinen malli**	**Karsittu malli**
MAL1	3.582	-
MAL2	38.895	-
MAL3	39.854	-
MKU1	6.571	-
MKU2	6.826	1.645
MKU3	4.476	2.802
MKU4	4.203	2.277
MPY1	1.043	1.043
MPY2	1.307	1.307
MPY3	1.350	1.350
MTE1	1.046	1.046
MTE2	1.680	1.680
MTE3	1.646	1.646
MUH1	1.094	1.094
MUH2	1.741	1.741
MUH3	1.707	1.707
MVA1	1.632	1.632
MVA2	1.875	1.875
MVA3	1.844	1.844
SOS1	1.003	1.003
SOS2	1.010	1.010
SOS3	1.011	1.011
VMO1	6.782	1.903
VMO2	7.534	-
VMO3	3.645	2.830
VMO4	2.580	2.257
VKÄ1	1.225	1.225
VKÄ2	1.225	1.225

6.2 Absoluuttiset ja suhteelliset vaikutukset

Muuttujan MKU (*mielletyt kustannukset*) indikaattorien vahvat ja tilastollisesti merkitsevät absoluuttiset kontribuutiot antavat viitteitä siitä, että indikaattorit mittaavat latenttia muuttujaa onnistuneesti. (ks. Taulukko 5). Konseptuaalista tärkeyttä mittaavat relatiiviset kontribuutiot ovat MKU2:n ja MKU 4:n osalta

myös tilastollisesti merkittävät, mutta MKU3:lle testi ei anna tukea. Miltei merkitsevän (p = 0,065) negatiivisen relatiivinen painoarvon vuoksi MKU3:n käyttäytymistä eri konteksteissa on kuitenkin syytä seurata.

Latentin muuttujan MPY (*minäpystyvyys*) indikaattoreista ainoastaan MPY1 saavutti tilastollisesti merkitsevät absoluuttiset ja suhteelliset painoarvot. Latenteista muuttujista MTE (*mielletty tehokkuus*), MUH (*mielletty uhkaavuus*) ja MVA (*mielletty vakavuus*) yksikään indikaattori ei osoittautunut tilastollisesti merkitseväksi niin absoluuttisilta kuin suhteellisilta vaikutuksiltaan. Cenfetellin ja Basselierin (2009) mukaan, indikaattorien poistamista malliin tulisi tässä tilanteessa harkita. Absoluuttiset vaikutusten tulokset kuitenkin antavat viitteitä muuttujien konseptuaalisen kentän heikosta taltioinnista. Tutkimuksen pienen otannan vuoksi ja muuttujien pienen indikaattorimäärän vuoksi todetaan, että muuttujat vaativat vielä tarkempaa tutkimista ja konseptuaalisen kentän perusteellista kartoittamista. MPY:n, MTE:n, MUH:n ja MVA:n epäonnistunut määrittely valitettavasti vaikuttaa muuttujien myöhempään analyysiin, kuten polkuanalyysiin. Koska muuttujien konseptuaalisten kenttien ei voida katsoa olevan edustavasti taltioitu, muuttujien polkuanalyysin painoarvot välttämismotivaatioon (VMO) vääristyvät (Petter ym., 2007).

Muuttujan SOS (*sosiaalinen vaikutus*) indikaattoreista SOS2 ja SOS3 saavat tukea absoluuttisten vaikutuksilleen ja SOS3 myös suhteellisille vaikutuksille. SOS1 osoittautui heikoimmaksi indikaattoriksi, eikä se saanut tukea absoluuttisille tai suhteellisille vaikutuksilleen. Sekä välttämismotivaation että välttämiskäyttäytymisen kaikki indikaattorit osoittavat mittaavan muuttujan varianssia onnistuneesti. Kuitenkin suhteellisilta vaikutuksiltaan merkitsevä näistä on vain VKÄ1. Tilastollisen merkitsevyyden rajan (p=0,05) ulkopuolelle jää täpärästi VMO1 ja loput indikaattorit, VMO3, VMO4 sekä VKÄ2 eivät saa tukea validiudelle.

Taulukko 5 Indikaattorien latautuminen latentteihin muuttujiin ja PLS-regression painoarvot

	Indikaattorien latautuminen (*absoluuttiset*)		**Regression painoarvot (*suhteelliset*)**	
Indikaattori → muuttuja	**Otoksen keskiarvo**	**p-arvo**	**Otoksen keskiarvo**	**p-arvo**
MKU2 → MKU	0.856	0.000	0.878	0.000
MKU3 → MKU	0.463	0.042	-0.536	0.065
MKU4 → MKU	0.657	0.000	0.631	0.011
MPY1 → MPY	0.872	0.000	0.842	0.000
MPY2 → MPY	-0.178	0.589	-0.239	0.486
MPY3 → MPY	0.205	0.330	0.138	0.584
MTE1 → MTE	0.398	0.175	0.423	0.175
MTE2 → MTE	0.053	0.373	-0.055	0.439
MTE3 → MTE	0.220	0.849	0.294	0.535
MUH1 → MUH	0.208	0.812	0.068	0.726

MUH2 → MUH	0.406	0.154	0.559	0.129
MUH3 → MUH	0.109	0.839	-0.236	0.196
MVA1 → MVA	-0.015	0.341	-0.142	0.308
MVA2 → MVA	0.191	0.887	0.265	0.259
MVA3 → MVA	0.097	0.348	0.002	0.292
SOS1 → SOS	-0.311	0.223	-0.289	0.258
SOS2 → SOS	0.594	0.046	0.542	0.066
SOS3 → SOS	0.597	0.015	0.526	0.035
VMO1 → VMO	0.873	0.000	0.573	0.055
VMO3 → VMO	0.826	0.000	0.252	0.643
VMO4 → VMO	0.762	0.000	0.246	0.122
VKÄ1 → VKÄ	0.932	0.000	0.853	0.000
VKÄ2 → VKÄ	0.557	0.021	0.199	0.436

6.3 Polkumuuttujat ja tutkimuksen hypoteesit

Muuttujien välisiä polkuja tutkiessa tilastollisesti merkitsevä painoarvoja löytyi vain kahdesta polusta (ks. Taulukko 6). Latenttien muuttujien MKU ja VMO (-0,494, p=0,002) välillä ilmeni kohtalaisen vahva negatiivinen painoarvo. MKU osoitti myös suurelta osin validiutta analyysin edellisessä vaiheessa, joten hypoteesi H4 saa tukea. Muuttujan VMO indikaattorit eivät osoittaneet luotettavuutta selittää yhdessä muuttujan varianssia. Vahvat absoluuttiset painoarvot ja tilastollisesti erittäin merkitsevä VMO-VKÄ -polun painoarvo (0,815, p=0,000) osoittaa, että latenttien muuttujien välillä vallitsee kuitenkin vahva yhteys. Tutkimuksen rajoitetun analyysiprosessin vuoksi H8 joudutaan kuitenkin hylkäämään. Empiirinen tuki osoittavat, että käyttäjän tiedon puute puolustusmenetelmistä sekä hänen kokemansa vaikeudet puolustusmenetelmien implementoimisessa ja riittävän suorituskykymetriikan tuloksen saavuttamisessa, ovat hyviä kandidaatteja kuvaamaan käyttäjän mieltämiä kustannuksia koneoppimisen kehitystehtävissä.

Polun SOS–VMO melko vahva painoarvo (0.395, p=0,150) ei ole tilastollisesti merkitsevä, eli tuen puutteessa H7 joudutaan hylkäämään. Samoin hypoteesit H2, H3, H5 ja H6 hylätään myös empiirisen tuen puuttumisen johdosta. Muuttujien välisiä vuorovaikutussuhteita tutkiessa tehtiin yleinen huomio, että konseptuaalisesti onnistuneemmin määritellyt muuttujat saivat polkuanalyysissä selkeästi vahvempia tuloksia, kuin epäonnistuneet muuttujat MPY, MTE, MUH, MVA. Tulokset tukevat Petterin ym. (2007) näkemystä, että muuttujien puutteellinen konseptuaalinen taltiointi johtaa haasteisiin analyysin kulussa ja tulosten tulkinnassa.

Taulukko 6 Polkuanalyysin tulokset ja tuet hypoteeseille

Polku	Hypoteesi	Otoksen keskiarvo	p-arvot	Tukeeko hypoteesia?
MKU → VMO	H4	-0.494	0.002	Kyllä
MPY → VMO	H6	0.072	0.912	Ei
MTE → VMO	H5	0.076	0.544	Ei
MUH → VMO	H3	0.155	0.123	Ei
MVA → MUH	H2	0.349	0.292	Ei
SOS → VMO	H7	0.395	0.150	Ei
VMO → VKÄ	H8	0.815	0.000	Kyllä

7 Pohdinta

Tässä luvussa käydään läpi tutkimustulosten implikaatiot teoreettisesta näkökulmasta. Alaluvussa 7.1 esitetään vastaukset tutkimuskysymyksiin ja alaluvussa 7.2 tulosten yleistettävyyteen ja luotettavuuteen vaikuttavat rajoitteet. 7.3 käsittelee tutkimuksen kontribuutioita sekä jatkotutkimusaiheita ja alalukuun 7.4 on koottu tutkielman johtopäätökset.

7.1 Tutkimuksen tulokset

Tutkielman alkuperäisenä päätavoitteena oli tutkia käyttäjien välttämiskäyttäytymistä suorituskykykriittisessä työympäristössä. Tätä ympäristöä edusti koneoppimisprosessi, jonka tuotteena syntyvän mallin arvon määrittää pitkälti sen kyky erotella havaintoja toisistaan halutulla tavalla. Tavoite pyrittiin toteuttamaan sovittamalla aiheeseen teoreettiseksi viitekehykseksi Teknologiauhkien välttämisen teoria (TTAT). Tutkimusprosessin kirjallisuuskatsauksen tukena hyödynnettiin kirjoittajan itse tekemää kandidaatintutkielmaa koneoppimisen sisältöluvun osalta. Tutkielmassa pyrkimyksenä oli selvittää vastaus tutkimuskysymykseen*:*

- *Mitkä tekijät vaikuttavat välttämiskäyttäytymiseen suorituskykykriittisessä työympäristössä?*

Valitettavasti tutkielma ei siis onnistunut tuottamaan luotettavaa vastausta tutkimuskysymykseen. Tutkimusprosessin edetessä ilmennyt olennainen informaatio muuttujien korrektista mallinnustavasta johti analyysiprosessin perustavanlaatuiseen muutokseen kyselytutkimuksen toteuttamisen jälkeen. Kyselytutkimusta laatiessa olisi pitänyt keskittyä mahdollisesti myös pienempään osaan mallista. Kyselyn koko pyrittiin pitämään suhteellisen pienenä, joten latenttien muuttujien lähtökohtaisesti suuri määrä karsi malliin sisällytettävien indikaattorien määrää. Reflektiivisesti mallinnetuille muuttujille tämä ei olisi aiheuttanut yhtä suurta ongelmaa, kuin tutkimuksen vaatimille formatiivisesti mallinnetuille muuttujille (Petter ym., 2007).

Kuudennessa luvussa suoritettu analyysi tuotti vaihtelevan laatuisia tuloksia. Latenttien muuttujien suurelta osin konseptuaalisesti kapea taltiointi vaikeutti analyysiprosessia Petterin (2007) osoittamaan tapaan ja teki tuloksista jokseenkin vaikeasti tulkittavia. Tutkimusaineiston analyysin pohjalta tukea löydettiin ainoastaan kahdelle hypoteesille:

H4: "Mielletyillä kustannuksilla on negatiiviset vaikutukset välttämismotivaatioon"

H8: "Välttämismotivaatiolla on positiivinen vaikutus välttämiskäyttäytymiseen"

Tuki miellettyjen kustannusten sisällyttämiselle poikkeaa Boysenin ym. (2019) tutkimuksen tuloksista. Vertailua aiempiin tutkimuksiin pitää kuitenkin tehdä varoen, sillä tutkielman formatiivisesti rakennettu malli poikkeaa rakenteellisesti muista tutkielmassa hyödynnetyistä reflektiivisistä malleista. Lisäksi mallin konseptuaaliset puutteet tarkoittavat, että myös polkuanalyysin painoarvojen luotettavuus joudutaan kyseenalaistamiseen (Petter ym., 2007).

Analyysissa ilmennyt tuen puute etenkin minäpystyvyyden, mielletyn tehokkuuden, mielletyn uhkaavuuden sekä mielletyn vakavuuden muuttujille antavat viitteitä kahdelle tulkinnalle: 1) Muuttujien konseptuaalinen kenttä ei olla taltioitu kattavasti, tai 2) Muuttujat eivät ole olennaisia mallissa. Bollen ja Lennox (1991) korostavat konseptuaalisen kentän kattavan taltioinnin erityisen suurta merkitystä formatiivisten muuttujien muodostamiselle. Absoluuttisten ja suhteellisten painoarvojen analyysi osoitti, että edellä mainittujen muuttujien indikaattorit eivät onnistu selittämään muuttujansa varianssia. Tämä näyttää heijastuvan myös heikkoihin painoarvojen merkitsevyyksiin polkuanalyysissa. Kirjallisuus ja analyysi johtavat siis päätelmään, että TTAT:n mukaisia uhkien sekä hallintakeinojen arviointiprossesseja ei onnistuttu taltioimaan malliin, jonka vuoksi polkumuuttujien painoarvot ja niiden suhteelliset vaikutukset välttämismotivaatioon vääristyvät. Tulosten heikkoon tilastolliseen merkittävyyteen vaikutti osaltaan kohderyhmän vaikea tavoitettavuus, mikä myös viestii kyselyn pieni 16:n vastauksen kokoinen otanta, jota ei voida pitää edustavana. Siten on myös huomioitava, että tämän tutkimuksen tuloksia ei voida laajentaa tutkimuksen otannan ulkopuolelle.

Tutkimuksen etenemisen aikana esiin nousseiden haasteiden ja kasvaneen ymmärryksen myötä TTAT:n teoreettisesta tutkimuksesta löytyi myös yksi epäkohta, jonka MacKenzie ym. (2005) ovat nostaneet tieteelliseen keskusteluun. Vallitseva käytäntö mallintaa latenttien muuttujien indikaattoreita reflektiivisesti perustelematta päätöstä ja huolimatta latentin muuttujan konseptuaalisesta moniulotteisuudesta voi pahimmassa tapauksessa johtaa puutteellisesti määriteltyihin teoreettisiin malleihin (Petter ym., 2007). Puutteelliselta määrittelyltä ei tosin säästytty myöskään tässä tutkielmassa. Tutkielman kirjoitusaikana ei löydetty TTAT-tutkimuksia, jotka hyödyntäisivät indikaattorien formatiivista mallintamista. Tässä tutkimuksessa pyritään myös nostamaan reflektiivisen mallintamisen käytäntö kriittisempään tarkasteluun TTAT:n tutkimuksen osalta ja priorisoida kattava konseptuaalinen mallintaminen ja teoreettinen kehitys tutkimuksen raportoitavia tuloksia korkeammalle.

Tutkimuskysymykseen vastaamisen epäonnistuttua, tutkimusprosessista voidaan sen sijaan nostaa muutamia jatkotutkimuksen tekijöitä hyödyttäviä tuloksia ja havaintoja. Koneoppimisen kontekstissa TTAT:n muuttujat ovat konseptuaalisesti moniulotteisia, ja siten niitä kuvaavia indikaattorit tulisi mallintaa formatiivisesti. Hullandin (1999) mukaan formatiiviset indikaattorit ovat aina riippuvaisia muista indikaattoreista, eli latentin muuttujan empiirinen merkitys muuttuu sitä ympäröivät kontekstin muuttuessa. Formatiivisessa mallinnuksessa muuttujien konseptuaalisesti kattavan taltioinnin suuri merkitys tulee siten

ottaa huomioon latenttien muuttujien indikaattoreita laatiessa (Howell ym., 2007).

7.2 Rajoitteet (Limitations)

Tutkielman kirjallisuuskatsauksessa analysoidun tieteellisen kirjallisuuden perusteella tämä tutkimus vaikuttaa olevan ensimmäinen, joka tutkii välttämiskäyttäytymistä koneoppimismallien kehittämisen kontekstissa. Teorian sovittaminen tähän uuteen tutkimushaaraan osoittautui hyvin haastavaksi, ja sen mukaisesti tällä tutkimuksella on myös useita rajoitteita. Selkeimpänä rajoitteena tutkimuksessa ovat mallin puutteellisuus. Koska aiempaa koneoppimista ja TTAT:ta yhdistävää tutkimusta ei ole, koneoppimisen aihealueelle spesifit indikaattorit ja kyselytutkimus jouduttiin rakentamaan itse. Pohjana indikaattorien määrittelylle käytettiin tutkielman kirjallisuuskatsauksen kolmatta lukua. Indikaattorien määrää jouduttiin tietoisesti rajaamaan suuremman otannan tavoittamisen toivossa. Useampi latentti muuttuja mallissa osoitti myös heikkoa validiutta, minkä vuoksi on oletettavaa, että latentit muuttujat eivät sisällä kaikkia konseptuaalisesti olennaisia indikaattoreita yleistettävien johtopäätösten tekemiseksi. Bollen ja Lennox (1991) korostavat, että muuttujien formatiivinen mallintaminen edustavasti vaatii, että indikaattorit kattavat muuttujan konseptuaalisen kentän kokonaisuudessaan. Jatkotutkimuksissa on siis tärkeää kartoittaa ja validoida koneoppimisen kontekstiin soveltuvat TTAT-indikaattorit.

Toisena merkittävänä puutteena toimii tutkielmassa useampaan kertaan mainittu otoskoon pieni määrä, joka myös heikentää tutkimustulosten yleistettävyyttä merkittävästi – tutkimuksen kohderyhmän tavoittaminen osoittautui maltillisista odotuksista huolimatta hyvin haastavaksi. Pieni otoskoko vaikutti negatiivisesti myös itse analyysiprosessiin. Bootstrapping-menetelmän hyödyntämisen mahdollistamiseksi jouduttiin kahdentamaan otos, joka on syytä ottaa huomioon analyysivaiheessa raportoituja tuloksia suhteellisista ja absoluuttisista vaikutuksista (Taulukko 5) sekä polkuanalyysin painoarvoista (Taulukko 6) tulkitessa. Tulokset edellä mainituista tilastollisista testeistä ovat tästäkin syystä parhaillaan vain suuntaa antavia.

7.3 Kontribuutiot ja jatkotutkimus

Ensimmäisenä ja suurimpana kontribuutiona tämä Pro Gradu -tutkielma nostaa tieteelliseen keskusteluun yhteiskunnallisesti hyvin tärkeän ja ajankohtaisen aihepiirin. Välttämiskäyttäytymistä koneoppimismallien kehityksessä – ja suorituskykykriittisissä työympäristöissä ylipäänsä – ei kirjoittajan parhaan tietämyksen mukaan olla tutkittu aiemmin tämän tutkielman palautusajankohtaan men-

nessä. Tutkielmaan kirjattu tutkimusprosessi ja sen haasteet tarjoavat siten tärkeää ymmärrystä aihepiiristä ja sen vaatimista erityishuomioista. Ollessaan polkunsa alkupäässä, tämä tutkimussuuntaus vaatii ehdottomasti jatkotutkimuksia. Erityistä huomiota tulee kiinnittää kohderyhmän tavoittamiseen ja formatiivisten muuttujien konseptuaalisen kentän kartoittamiseen. Tutkielma tarjoaa lähtöpisteen koneoppimisen hyökkäysalan huomioivan tutkimuksen tekemiselle: jatkotutkimuksen tekijä voi hyödyntää kyselytutkimusta pohjana, jota jatkokehittää, sekä laajentaa ja testata tutkielmaa varten luotua aihepiirille spesifiä indikaattorijoukkoa.

Toisena kontribuutiona tutkielman kirjallisuuskatsaus kokoaa koneoppimisen tutkimuksen aikana kumuloitunutta ymmärrystä. Erityisen oleellinen osio kirjallisuuskatsauksessa on tietojärjestelmätieteen tutkimuksessa vähäiselle huomiolle jäänyt koneoppimisen vahingolliseen IT:aan verrattava kääntöpuoli, vihamielinen koneoppiminen, josta ei varsinkaan suomen kielellä löydy riittävästi tietoa. Alalukuun 3.4 on koottu tietoa koneoppimisen hyökkäysalasta sekä vihamielisestä koneoppimisesta, jonka avulla tutkijat ja lukijat voivat saada aihepiiristä realistisemman yleiskuvan.

Kolmantena ja ehkäpä olennaisimpana kontribuutiona tämä tutkielma osoittaa tärkeän kehityssuunnan TTAT-tutkimuksessa latenttien muuttujien ja indikaattorien välisten suhteiden mallintamisessa. Reflektiiviset indikaattorit pelkästään eivät kykene kuvaamaan riittävän kattavasti koneoppimisen moniulotteisia ja monimutkaisia vuorovaikutussuhteita. Indikaattorien formatiivinen mallintaminen on välttämiskäyttäytymisen tutkimuksessa hyvin olennainen kehityssuunta, kun tutkittavat muuttujat toimivat osana kompleksista prosessia, kuten koneoppimismallien kehitystehtävissä tapahtuvaa käyttäjän päätöksentekoprosessia. Tietojärjestelmätieteen tutkimuksessa käytetyissä latenttien muuttujien mallinnustavoissa on ilmennyt lukuisia puutteita (Petter ym., 2007), jotka on hyvä korjata TTAT:n ollessa vielä suhteellisen varhaisessa vaiheessa teoreettista kehitystään. Tulevissa TTAT-tutkimuksissa on siis syytä kiinnittää huomiota latenttien muuttujien konseptuaalisen kentän mallintamiseen edustavasti ja kokonaisvaltaisesti. Koska formatiiviset indikaattorit ovat täysin riippuvaisia ympäröivästä kontekstista (Howell ym., 2007), on epäselvää, mitä indikaattoreita latenttien muuttujien tulisi tarkalleen ottaen sisältää, ja kuinka hyvin tähän tutkimukseen sisällytetyt indikaattorit soveltuvat yhteen eri konteksteissa – joka tapauksessa tutkielman malli vaatii kipeästi jatkokehitystä. Indikaattorien validoinnilla on siis suuri merkitys koneoppimismallien kehitystehtäviin liittyvän välttämiskäyttäytymisen tutkimuksen kannalta.

7.4 Johtopäätökset

Tämän Pro Gradu -tutkielman tarkoituksena oli selvittää, miten käyttäjän pyrkivät välttämään uhkia suorituskykykriittisessä ympäristössä. Tutkimuksen aihepiiriksi valittiin koneoppimismallien kehitysprosessi, jota on tieteellisessä tutki-

muksessa käsitelty kirjallisuuskatsauksen perusteella verrattain vähän turvallisuusnäkökulmasta. Koneoppimisen yhdistäminen Teknologiuhkien välttämisen teoriaan (TTAT) on tieteellisessä tutkimuksessa täysin uusi suuntaus.

Tutkielma muodostettiin kahdesta pääosiosta, kirjallisuuskatsauksesta ja empiirisestä osiosta. Kirjallisuuskatsaus jaettiin teorian ja aihepiirin kesken TTAT:ta (luku 2) ja koneoppimista (luku 3) käsitteleviin lukuihin, joiden päätteeksi koottiin kirjallisuuskatsauksen tärkeimmät huomiot. Empiirinen osio suoritettiin strukturoituna survey-tutkimuksena ja kerätty aineisto analysoitiin kvantitatiivisin menetelmin. Neljännessä luvussa kerrattiin tutkimusongelma ja esiteltiin hypoteesit sekä kirjallisuuskatsauksen pohjalta muodostettu malli. Viidennessä luvussa käytiin läpi tutkimuksen analyysiprosessi, ja kuudenteen lukuun koottiin analyysiprosessin tuottamat tulokset. Tutkielman viimeinen, seitsemänteen lukuun kirjattiin tutkimustulokset, tutkimuksen rajoitteet sekä johtopäätökset.

Varsinaiseen tutkimuskysymykseen tutkielma ei kyennyt tuottamaan vastausta. Tutkimusympäristön tuntemattomuus teorian näkökulmasta nosti tutkielman etenemisen aikana kasvavan tarpeen tutkimussuunnan itsenäiselle kartoittamiselle. Tutkimusprosessin aikana ilmeni lukuisia haasteita, joiden yhteisvaikutukset tutkielman kirjoittamisen aikapaineen ja tutkijan kokemattomuuden kanssa johtivat puutteisiin, johtaen epäonnistumiseen tutkimuskysymyksen vastaamisen osalta.

Edellä mainitusta epäonnistumisesta huolimatta tutkielma avaa väylän hyvin tärkeän ja uuden tutkimussuunnan koneoppimisen välttämiskäyttäytymisen tutkimukselle. Tutkimuksen tuloksena tuotettiin aihepiiriin kattavasti esittelevä kirjallisuuskatsaus, koneoppimisen hyökkäysalan huomioiva kyselytutkimus, sekä reilu määrä koneoppimiselle spesifejä indikaattoreita TTAT-jatkotutkimuksia varten.

Koneoppimismallien kehittämiseen liittyvä problematiikka on hyvin moniulotteinen: kun samoja menetelmiä voidaan hyödyntää niin koneoppimismallin kehittämiseen kuin sen vahingoittamiseenkin, on mahdotonta löytää yksiselitteisiä keinoja, joilla koneoppimisprosessien turvallisuus voidaan taata. Uusien menetelmien kehittäminen ja ymmärryksen syventäminen olemassa olevista menetelmistä, tuottavat arvokasta tietoa koneoppimisen mahdollisuuksista niin hyveellisissä kuin vahingollisissakin tarkoituksissa.

Välttämiskäyttäytymisen tutkiminen koneoppimisen kehittämisen kontekstissa on tarpeellinen tutkimussuunta niin TTAT:n teoreettisen kehityksen kuin koneoppimisenkin kannalta. Hyökkäyksille resilienttien menetelmien kehittäminen on hyvin olennainen osa koneoppimisprosessin turvaamista tulevaisuudessa, mutta sitäkin suurempaan asemaan nousee käyttäjien motivaatio ja aktiiviset pyrkimykset välttää koneoppimiseen kohdistuvia uhkia jokapäiväisessä työssään. Välttämismotivaatiota lisäävien ja heikentävien tekijöiden identifioiminen on tärkeä askel koneoppimisen uhkatekijät huomioon ottavan ajatusmaailman rakentamisessa käyttäjien keskuuteen. Organisaatioiden päätöksenteon nojatessa yhä enenevissä määrin koneoppimismenetelmien avulla tuotettuun infor-

maatioon, käyttäjien rooli organisaatioiden liiketoiminnan jatkuvuuden edistäjänä tulee kasvamaan yhä edelleen. Koneoppimisen kehittäjien välttämiskäyttäytymisellä tulee olemaan suuri merkitys siinä, miten haavoittuvaisia organisaatiot ovat tulevaisuuden kybertoimintaympäristöissä.

LÄHTEET

Abadi, M., McMahan, H. B., Chu, A., Mironov, I., Zhang, L., Goodfellow, I., & Talwar, K. (2016a). Deep learning with differential privacy. *Proceedings of the 23rd ACM Conference on Computer and Communications Security,* 308–318.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Zheng, X. (2016b). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, 265–283.

Abbasi, M. & Gagné, C. (2017). *Robustness to Adversarial Examples Through an Ensemble of Specialists.* ArXiv.

Alpaydin, E. (2010). *Introduction to Machine Learning* (2. painos). London, England: The MIT Press.

Alpaydin, E. (2016). *Machine Learning: The New AI* (3. painos). London, England: MIT Press.

Amazon Web Services. *Machine Learning on AWS*. Haettu 11.1.2020 osoitteesta https://aws.amazon.com/machine-learning/

Aon (2019). *2019 Cyber Security Risk Report*. Haettu 16.12.2019 osoitteesta https://www.aon.com/unitedkingdom/insights/2019-cyber-security-risk-report.jsp

Arachchilage, N. A. G., & Love, S. (2013). A Game Design Framework for Avoiding Phishing Attacks. *Computers in Human Behavior, 29(3),* 706-714.

Aslett, L. J. M., Esperança, P. M. & Holmes, C. C. (2015). *Encrypted Statistical Machine Learning: New Privacy Preserving Methods.* ArXiv.

Bagchi, K., and Udo, G. 2003. An Analysis of the Growth of Computer and Internet Security Breaches. *Communications of the AIS*, 684-700.

Bai, E. W. (2014). Big data: The Curse of Dimensionality in Modeling. *Proceedings of the 33rd Chinese Control Conference*, 6–13.

Barreno, M., Nelson, B., Joseph, A. D. & Tygar, J. D. (2010). The Security of Machine Learning. *Machine Learning, 81(2),* 121–148.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D. & Tygar, J. D. (2006). Can Machine Learning Be Secure? *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security,* 16-25.

Barsky, R. B., Juster, F. T., Kimball, M. S. & Shapiro, M. D. (1997). Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study. *The Quarterly Journal of Economics, 72(3),* 537-579.

Barto, A. & Dietterich, T. (2004). Reinforcement learning and its relationship to supervised learning. *Handbook of Learning and Approximate Dynamic Programming,* 47–64.

Biggio, B., Fumera, G., & Roli, F. (2014). Pattern Recognition Systems under Attack: Design Issues and Research Challenges. *International Journal of Pattern Recognition and Artificial Intelligence, 28(7),* 1–22.

Bollen, K. & Lennox, R. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin, 110(2),* 305–314.

Boysen, S., Hewitt, B., Gibbs, D., & McLeod, A. (2019). Refining the Threat Calculus of Technology Threat Avoidance Theory. *Communications of the Association for Information Systems*, 45(5).

Buczak, A., & Guven, E. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials, 18(2),* 1153-1175.

Carpenter, D., Young, D.K., Barrett, P., & McLeod, A.J. (2019). Refining Technology Threat Avoidance Theory. *Communications of the Association for Information Systems, 44(22),* 380 – 407.

Carver, C. S. (2006). Approach, Avoidance, and the Self-Regulation of Affect and Action. *Motivation and Emotion, 30(2)*, 105-110.

Carver, C. S. & Scheier, M. F. (1982). Control Theory: A Useful Conceptual Framework for Personality-Social, Clinical, and Health Psychology. *Psychological Bulletin, 92(1),* 111-135.

Cenfetelli, R. T. & Bassellier, G. (2009). Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly, 33(4),* 689-707

Corona, I., Giacinto, G., & Roli, F. (2013). Adversarial Attacks Against Intrusion Detection Systems: Taxonomy, Solutions, and Open issues. *Information Sciences, 239*, 201–225.

CyberEdge Group (2019). 2019 Cyberthreat Defense Report.

Deutsch, M., and Gerard, H. B. (1955). A Study of Normative and Informational Social Influences upon Individual Judgment. *Journal of Abnormal and Social Psychology (51),* 629-636.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research, 38(2),* 269-277.

Diamantopoulos, A. & Siguaw, J. A. 2006. Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management, 17(4),* 263-282.

Eberhart, R., Simpson, P. & Dobbins, R. (1996). *Computational Intelligence PC Tools.* Orlando, Florida: Academic Press.

Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, Florida: Chapman & Hall/CRC.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. 2015 International Conference on Learning Representations, 1–11. ArXiv.

Feng, C., Wu, S., & Liu, N. (2017). A User-centric Machine Learning Framework for Cyber Security Operations Center. IEEE International Conference on Intelligence and Security Informatics, 173–175.

Fraley, J. B. & Cannady, J. (2017). The Promise of Machine Learning in Cybersecurity. Conference Proceedings - IEEE SOUTHEASTCON.

Hastie, T., Tibshirani, R. & Friedman. J. (2011). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second edition). New York, USA: Springer.

Howell, R., Breivik, E. & Wilcox, J.B. (2007). Reconsidering Formative Measurement. Psychological Methods, 12(2), 205–218.

Hulland, J. (1999). Use of Partial Least Squares (PLS) in Strategic Management Research: A Review of Four Recent Studies. Strategic Management Journal, 20(2), 195- 204

IBM Watson. IBM Watson Products and Solutions. Haettu 12.1.2020 osoitteesta https://www.ibm.com/watson/

ISO/IEC 27000:2018. *Information technology – Security techniques – Information Security Management Systems – Overview and Vocabulary.* Haettu 5.1.2020 osoitteesta http://isotc.iso.org/livelink/livelink/fetch/2000/ 2489/ Ittf_Home/PubliclyAvailableStandards.htm

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. *Proceedings – 2018 IEEE Symposium on Security and Privacy*, 19–35.

Kim, K., Erza, M., Harry, A., & Tanuwidjaja, C. (2018). Network Intrusion Detection using Deep Learning: A Feature Learning Approach. *SpringerBriefs on Cyber Security Systems and Networks.* Singapore: Springer.

Kloft, M. & Laskov, P. (2010). Online Anomaly Detection Under Adversarial Impact. *International Conference on Artificial Intelligence and Statistics,* 405–412.

Lazarus, R. (1966). *Psychological Stress and the Coping Process.* New York: McGraw-Hill.

Lazarus, R. & Folkman, S. (1984). *Stress, Coping, and Adaptation.* New York: Springer-Verlag.

Lehto, M., Limnéll, H., Innola, E., Pöyhönen, J., Rusi, T. & Salminen, M. (2017). Suomen kyberturvallisuuden nykytila, tavoitetila ja tarvittavat toimenpiteet tavoitetilan saavuttamiseksi. *Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 30/2017*. Valtioneuvoston kanslia.

Liang, H. & Xue, Y. (2009). Avoidance of Information Technology Threats: A Theoretical Perspective. *MIS quarterly 33(1),* 71-90.

Liang, H. & Xue, Y. (2010). Understanding Security Behaviors in Personal Computer Usage: A Threat Avoidance Perspective. *Journal of the Association for Information Systems, 11(7),* 394-413.

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. M. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access, 6*, 12103–12117.

MacKenzie, S.B., Podsakoff, P.M., & Jarvis, C.B. (2005). The Problem of Measurement Model Misspecification in Behavioral and Organizational Research and Some Recommended Solutions. *Journal of Applied Psychology, 90(4)*, 710–730.

Manadhata, P. K., & Wing, J. M. (2011). An attack surface metric. *IEEE Transactions on Software Engineering, 37(3),* 371–386.

Michalski, R., Carbonell, J. & Mitchell, T. (1985). Machine learning: An Artificial Intelligence Approach. *Artificial Intelligence, 25(2),* 236–238.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I. P., Saini, U., … Xia, K. (2008). Exploiting Machine Learning to Subvert Your Spam Filter. *Proceedings of the First Workshop on Large-Scale Exploits and Emerging Threats.*

Papernot, N., McDaniel, P. & Goodfellow, I. (2016a). *Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples.* ArXiv.

Papernot, N., McDaniel, P., Sinha, A. & Wellman, M. (2016b*). SoK: Towards the Science of Security and Privacy in Machine Learning.* ArXiv.

Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016c). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *Proceedings of 2016 IEEE Symposium on Security and Privacy*, 582–597.

Petter, S., Straub, D. & Rai, A. (2007). Specifying Formative Construicts in IS Research. *MIS Quarterly, 31(4),* 657-679.

Sanastokeskus TSK. (2018). *Kyberturvallisuuden sanasto (TSK 52).* Huoltovarmuuskeskus, Helsinki 2018.

Sapp, C. E. (2018). *Preparing and Architecting for Machine Learning.* Gartner Technical Professional Advice.

Sitkin, S. B., & Weingart, L. R. (1995). Determinants of Risky Decision-Making Behavior: A Test of the Mediating Role of Risk Perceptions and Propensity. *Academy of Management Journal, 38(6),* 1573-1592.

SmartPLS (2020*). SmartPLS (ohjelmisto), versio 3.3.2.* https://www.smartpls.com/

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing Properties of Neural Networks. *2nd International Conference on Learning Representations*, 1–10.

Tramèr, F., Kurakin, A., Papernot, N. N., Goodfellow, I., Boneh, D. & McDaniel P. (2018). *Ensemble Adversarial Training: Attacks and defenses.* ArXiv.

Tversky, A., and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science 185(4157),* 1124-1131.

Vinzi, V. E., Trinchera, L., & Amato, S. (2010). *PLS Path modeling: from Foundations to Recent Developments and Open Issues for Model Assessment and Improvement,* 47-82. Berlin, Heidelberg: Springer.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3. painos.). The Morgan Kaufmann Series in Data Management Systems.

Yao, Y.C. Song L. & Chi, E. (2017). Investigation on Distributed K-means Clustering Algorithm of Homomorphic Encryption. *Computational Technology Development,* 81–85.

Young, D. K., Carpenter, D., & McLeod, A. (2016). Malware avoidance motivations and behaviors: A Technology Threat Avoidance Replication. *AIS Transactions on Replication Research, 2(8),* 1-17.

Zanero, S. & Serazzi, G. (2008). Unsupervised Learning Algorithms for Intrusion Detection. *Network Operations and Management Symposium* 2008, 1043–1048.

LIITTEET

Master´s thesis questionnaire

INTRODUCTION

This Master's Thesis research studies users' threat perceptions and avoidance behavior in the context of Machine Learning development. This survey questionnaire is targeted to people who utilize Machine Learning in their work. This survey, and the thesis including it, are a part of my Master's degree at the University of Jyväskylä, and are not affiliated with any third party organizations.

This survey questionnaire consists of 30 questions and takes approximately 9–12 minutes to complete.

PURPOSE OF THE SURVEY

This survey's purpose is to examine and achieve a better understanding on *user threat avoidance behavior* with respect to the Technology Threat Avoidance Theory (TTAT), and in the context of Machine Learning development.
This survey does not attempt in any way to take a stand on a "correct" way to develop Machine Learning models.
The respondents of this survey should answer the questions according to their personal views, perceptions and beliefs.

IMPORTANT TERMS IN THE CONTEXT OF THIS SURVEY

This survey includes two important terms, that are formed for the purpose of this survey, and exist to simplify the formulation and phrasing of the survey questions. These two terms are referred to several times within the survey, so their meaning should be remembered. The questions will have these terms underlined to remind you of their special meaning in the context of this survey. The terms are defined below.

An attack refers to an adversary's attempt to inflict negative effects on a Machine Learning model, the information produced during the learning process, or to steal information concerning the model's functionality. These attacks may occur during the Machine Learning model's training or inference phase. Examples of potential attacks are:

1. Manipulations made to a Machine Learning model's training data or hyperparameters to undermine the model's intended purpose (*e.g. data poisoning, logic corruption*)
2. Extracting information from the model and its training data to copy the model's functionality (i.e. *surrogate models*)
3. Revealing and exploiting vulnerabilities (*i.e. misclassified observations, inaccuracies in decision boundary*) by querying the model

Defensive techniques refer to the Developer's / Data Scientist's attempt to reduce the risk of a successful attack and mitigate their effects, and to ensure the model's generalizability. Examples of defensive techniques are:

1. Exclusion of adversarial inputs from training data (i.e. *data sanitization*)
2. Learning information about the adversary's attacks and attempting to utilize the extracted information to train robust and resilient models (i.e. *adversarial training*)
3. Other techniques applied / performed **with the intention to protect** the Machine Learning model, its training data, and the information generated, from manipulations (e.g. *information hiding*)

Feel free to contact me for feedback or any questions you might have:

Topi Luukkanen
Master's degree student (Information Systems Science)
University of Jyväskylä
tomaluuk@student.jyu.fi

BACKGROUND INFORMATION

1. Age *

- ◯ < 20
- ◯ 20-24
- ◯ 25-29
- ◯ 30-34
- ◯ 35-39
- ◯ 40-44
- ◯ > 45

2. Experience with Machine Learning *

- ◯ Very little / no experience
- ◯ 1-3 years
- ◯ 4-6 years
- ◯ 7-9 years
- ◯ 10+ years

Next you will answer questions and claims about various security aspects of Machine Learning. Choose the option that best reflects Your own personal views, perceptions and beliefs – there are no wrong answers.

The answer options are:

Strongly disagree
Disagree
Somewhat disagree
Undecided
Somewhat agree
Agree
Strongly agree.

3. It is extremely likely that my Machine Learning models, their data, or the outputs produced by them, will be attacked by an adversary in the future *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

4. There is a good possibility that my Machine Learning models, their data, or the outputs produced by them, will be attacked by an adversary at some point. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

5. There is a good chance that my Machine Learning models, their data, or the outputs produced by them, will be attacked by an adversary at some point in the future. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

6. Observations in my Machine Learning model's training data could be added, deleted or perturbed without my knowledge. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

7. The information collected from my Machine Learning models could be misused by an adversary. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

8. An adversary's attack could deteriorate the generalization performance of my Machine Learning models, preventing their intended use. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

9. An adversary's attack targeted to my Machine Learning model poses a threat to me. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

10. It would be dreadful if the integrity of information in my Machine Learning model was violated by an adversary's attack. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

11. It would be risky to use my Machine Learning model if it had been attacked by an adversary. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

12. Defensive techniques would increase my ability to protect my Machine Learning models from an adversary's attack. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

13. Defensive techniques would enable me to train Machine Learning models, with sufficient generalization performance, faster. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

14. Defensive techniques would increase my productivity in training sufficiently performing Machine Learning models under the presence of an adversary. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

15. I don't utilize defensive techniques on my Machine Learning models because I am not familiar with them. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

16. I don't utilize defensive techniques on my Machine Learning models, because I'm not familiar with the aspects that need to be considered in their use. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

17. I don't utilize defensive techniques on my Machine Learning models because they complicate the learning process. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

18. I don't utilize defensive techniques on my Machine Learning models because it's more difficult to achieve a high enough performance metric score that is expected from me (e.g. recall, accuracy, F1 score). *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

19. My superiors expect that my Machine Learning models should achieve a high score on chosen performance metric (e.g. recall, accuracy, F1 score) *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

20. My peers think that I should utilize defensive Machine Learning techniques to protect my models from attacks. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

21. In general, people have supported/recommended the use of defensive Machine Learning techniques. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

22. I could successfully implement defensive Machine Learning techniques to my Machine Learning models even if there was no one around to tell me what to do. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

23. I could implement defensive Machine Learning techniques to my models if I had a lot of time to complete the task. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	◯	◯	◯	◯	◯	◯	◯

24. I could implement defensive Machine Learning techniques to my models if someone showed me how to do it first. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

25. I intend to use defensive Machine Learning techniques to avoid adversary's attacks to my models. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

26. I plan to use defensive Machine Learning techniques to avoid adversary's attacks to my models. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

27. I research information about defensive Machine Learning techniques regularly. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

28. I research information about the vulnerabilities of Machine Learning regularly. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

29. I use defensive Machine Learning techniques to identify potential attacks and mitigate their effects on my models. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

30. I avoid utilizing Machine Learning techniques that have been found prone to attacks. *

	Strongly disagree	Disagree	Somewhat disagree	Undecided	Somewhat agree	Agree	Strongly agree
Your answer	○	○	○	○	○	○	○

You have completed the questionnaire. Please remember to submit your answers!

I thank you for your time and answers. If you wish to give feedback about the survey, you can do so in the answer field below. You can also contact me directly via email.

If you are interested in the subject, here's a couple interesting articles about the topic:

Papernot, N., McDaniel, P., Sinha, A. & Wellman, M. (2016b). SoK: Towards the Science of Security and Privacy in Machine Learning. ArXiv. (https://arxiv.org/pdf/1611.03814.pdf)

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. M. (2018). A survey on security threats and defensive techniques of machine learning: A data driven view. IEEE Access, 6, 12103–12117. (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8290925)

Liang, H. & Xue, Y. (2010). Understanding security behaviors in personal computer usage: A threat avoidance perspective. Journal of the Association for Information Systems, 11(7), 394-413. (https://pdfs.semanticscholar.org/6685/906db2746723ff00a377e66a391235020314.pdf)

Kind regards,

Topi Luukkanen
Master's degree student (Information Systems Science)
University of Jyväskylä
tomaluuk@student.jyu.fi

31. Do you have feedback?

Tilastollinen kuvailu kaksinkertaistetusta datasta

	Numero	Puuttuvia	Keskiarvo	Mediaani	Keskihajonta	Huipukkuus	Vinous
MAL1	1	0	3.875	4	1.728	-1.255	-0.255
MAL2	2	0	4.500	5	1.458	0.267	-1.080
MAL3	3	0	4.562	5	1.499	0.150	-1.062
MVA1	4	0	3.812	4	1.629	-1.399	-0.137
MVA2	5	0	4.625	5	1.615	-1.088	-0.563
MVA3	6	0	4.625	5	1.576	-0.938	-0.544
MUH1	7	0	3.750	5	1.479	-1.789	-0.152
MUH2	8	0	5.062	6	1.345	-0.058	-0.929
MUH3	9	0	4.938	5	1.519	-0.892	-0.338
MTE1	10	0	5.625	6	0.781	-0.310	-0.026
MTE2	11	0	3.812	4	1.590	-1.233	-0.166
MTE3	12	0	4.375	4	1.452	-0.003	-0.703
MKU1	13	0	4.312	5	1.722	-0.914	-0.514
MKU2	14	0	4.438	5	1.413	0.416	-1.123
MKU3	15	0	3.875	4	1.409	-0.664	-0.330
MKU4	16	0	3.312	4	1.402	-0.732	-0.023
SOS1	17	0	3.312	3	1.446	-1.155	0.329
SOS2	18	0	3.750	4	1.346	-1.142	0.161
SOS3	19	0	3.562	4	1.580	-0.355	0.478
MPY1	20	0	4.500	5	1.369	-0.429	-0.383
MPY2	21	0	5.375	5	1.111	-1.258	0.332
MPY3	22	0	6.062	6	0.899	-0.329	-0.670
VMO1	23	0	4.750	5	1.090	1.341	-0.380
VMO2	24	0	4.625	5	1.166	0.437	-0.194
VMO3	25	0	3.688	5	2.083	-1.584	-0.039
VMO4	26	0	4.312	5	1.927	-1.197	-0.524
VKÄ1	27	0	2.938	2	1.819	-0.323	0.948
VKÄ2	28	0	4.125	5	1.452	-1.279	-0.357

Indikaattorien latautuminen latentteihin muuttujiin

	Alkuperäinen otos	Otoksen keskiarvo	Keskihajonta	Studentin t-jakauma	p-arvot
MKU2 → MKU	0.885	0.856	0.151	5.854	0.000
MKU3 → MKU	0.521	0.463	0.256	2.037	0.042
MKU4 → MKU	0.726	0.657	0.201	3.605	0.000
MPY1 → MPY	0.972	0.872	0.209	4.645	0.000
MPY2 → MPY	-0.175	-0.178	0.324	0.540	0.589
MPY3 → MPY	0.224	0.205	0.229	0.975	0.330
MTE1 → MTE	0.935	0.398	0.690	1.355	0.175
MTE2 → MTE	-0.425	0.053	0.477	0.891	0.373
MTE3 → MTE	-0.080	0.220	0.421	0.190	0.849
MUH1 → MUH	0.117	0.208	0.493	0.238	0.812
MUH2 → MUH	0.723	0.406	0.507	1.426	0.154
MUH3 → MUH	-0.085	0.109	0.417	0.203	0.839
MVA1 → MVA	-0.547	-0.015	0.575	0.953	0.341
MVA2 → MVA	0.066	0.191	0.468	0.142	0.887
MVA3 → MVA	-0.645	0.097	0.687	0.939	0.348
SOS1 → SOS	-0.427	-0.311	0.350	1.218	0.223
SOS2 → SOS	0.705	0.594	0.353	1.994	0.046
SOS3 → SOS	0.662	0.597	0.273	2.426	0.015
VMO1 → VMO	0.929	0.873	0.169	5.499	0.000
VMO3 → VMO	0.844	0.826	0.162	5.202	0.000
VMO4 → VMO	0.824	0.762	0.149	5.530	0.000
VKÄ1 → VKÄ	0.976	0.932	0.138	7.063	0.000
VKÄ2 → VKÄ	0.615	0.557	0.267	2.305	0.021

Indikaattorien osittaiset painoarvot PLS-regression jälkeen

	Alkuperäi-nen otos	Otoksen keskiarvo	Keskiha-jonta	Studentin t-jakauma	p-arvot
MKU2 → MKU	0.881	0.878	0.189	4.665	0.000
MKU3 → MKU	-0.552	-0.536	0.299	1.845	0.065
MKU4 → MKU	0.699	0.631	0.274	2.551	0.011
MPY1 → MPY	0.940	0.842	0.229	4.100	0.000
MPY2 → MPY	-0.261	-0.239	0.375	0.697	0.486
MPY3 → MPY	0.181	0.138	0.331	0.548	0.584
MTE1 → MTE	0.893	0.423	0.658	1.356	0.175
MTE2 → MTE	-0.451	-0.055	0.584	0.773	0.439
MTE3 → MTE	0.338	0.294	0.545	0.620	0.535
MUH1 → MUH	-0.180	0.068	0.512	0.351	0.726
MUH2 → MUH	1.307	0.559	0.860	1.520	0.129
MUH3 → MUH	-0.903	-0.236	0.697	1.295	0.196
MVA1 → MVA	-0.584	-0.142	0.574	1.019	0.308
MVA2 → MVA	0.999	0.265	0.886	1.128	0.259
MVA3 → MVA	-0.951	0.002	0.903	1.054	0.292
SOS1 → SOS	-0.383	-0.289	0.338	1.131	0.258
SOS2 → SOS	0.640	0.542	0.348	1.840	0.066
SOS3 → SOS	0.582	0.526	0.276	2.111	0.035
VMO1 → VMO	0.617	0.573	0.322	1.916	0.055
VMO3 → VMO	0.167	0.252	0.360	0.464	0.643
VMO4 → VMO	0.347	0.246	0.224	1.546	0.122
VKÄ1 → VKÄ	0.873	0.853	0.208	4.199	0.000
VKÄ2 → VKÄ	0.241	0.199	0.309	0.780	0.436

Polkuanalyysin tulokset

Polku	Alkuperäinen otos	Otoksen keskiarvo	Keskihajonta	Studentin t-jakauma	p-arvot
MTE -> VMO	-0.118	0.075	0.194	0.605	0.545
MUH -> VMO	0.426	0.154	0.283	1.504	0.133
MVA -> MUH	0.757	0.350	0.716	1.057	0.291
MKU -> VMO	-0.670	-0.494	0.216	3.093	0.002
MPY -> VMO	-0.023	0.073	0.210	0.112	0.911
SOS -> VMO	0.456	0.392	0.319	1.431	0.152
VMO -> VKÄ	0.795	0.815	0.066	12.027	0.000

	Ikä	Kokemus	MAL1	MAL2	MAL3	MVA1	MVA2	MVA3	MUH1	MUH2	MUH3	MTE1	MTE2	MTE3
Ikä	1	0.38	0.55	0.58	0.54	0.03	0.14	0.09	0.46	0.23	-0.06	-0.14	0.03	-0.07
Kokemus	0.38	1.00	0.01	0.2	0.19	-0.13	-0.05	-0.12	0.26	0.13	-0.07	0.21	-0.02	0.17
MAL1	0.55	0.01	1	0.84	0.85	0.55	0.34	0.28	0.48	0.19	0.21	-0.50	0.26	0.17
MAL2	0.58	0.20	0.84	1	0.99	0.33	0.32	0.33	0.72	0.05	-0.01	-0.44	0.23	0.24
MAL3	0.54	0.19	0.85	0.99	1	0.35	0.37	0.35	0.74	0.08	0.04	-0.46	0.18	0.25
MVA1	0.03	-0.13	0.55	0.33	0.35	1	0.57	0.56	0.21	-0.31	-0.03	-0.45	0.42	0.21
MVA2	0.14	-0.05	0.34	0.32	0.37	0.57	1	0.63	0.61	0.01	-0.16	-0.31	0.36	0.3
MVA3	0.09	-0.12	0.28	0.33	0.35	0.56	0.63	1	0.44	-0.28	0.04	-0.17	0.50	0.25
MUH1	0.46	0.26	0.48	0.72	0.74	0.21	0.61	0.44	1	0.16	-0.09	-0.14	0.06	0.28
MUH2	0.23	0.13	0.19	0.05	0.08	-0.31	0.01	-0.28	0.16	1	0.61	0.44	-0.14	0.18
MUH3	-0.06	-0.07	0.21	-0.01	0.04	-0.03	-0.16	0.04	-0.09	0.61	1	0.19	-0.19	0.1
MTE1	-0.14	0.21	-0.50	-0.44	-0.46	-0.45	-0.31	-0.17	-0.14	0.44	0.19	1	-0.21	-0.15
MTE2	0.03	-0.02	0.26	0.23	0.18	0.42	0.36	0.50	0.06	-0.14	-0.19	-0.21	1	0.63
MTE3	-0.07	0.17	0.17	0.24	0.25	0.21	0.3	0.25	0.28	0.18	0.1	-0.15	0.63	1
MKU1	-0.48	-0.44	0.18	-0.06	-0.04	0.47	0.2	-0.14	-0.14	-0.09	-0.09	-0.42	0.25	0.28
MKU2	-0.43	-0.41	0.05	-0.14	-0.12	0.23	0.02	-0.18	-0.1	0.02	-0.05	-0.19	0.12	0.35
MKU3	0.04	-0.45	0.2	-0.03	-0.06	0.26	0.14	0.12	0.07	0.3	0.23	0.01	0.27	0.24
MKU4	0.13	-0.37	0.45	0.08	0.07	0.49	0.41	0.11	0.01	0.25	0.07	-0.24	0.42	0.1
SOS1	-0.48	-0.14	-0.19	-0.01	0.02	-0.16	-0.37	-0.02	-0.21	-0.05	0.25	0.06	0.06	0.29
SOS2	0.13	-0.02	0.01	-0.1	-0.12	-0.31	-0.33	-0.25	-0.28	0.39	0.63	-0.03	-0.14	-0.08
SOS3	0.09	0.18	0.05	0.12	0.08	-0.32	-0.36	-0.14	-0.07	0.25	0.12	0.53	-0.11	-0.42
MPY1	0.52	0.39	0	0.13	0.14	-0.32	0.11	0.12	0.28	0.22	0.02	0.23	-0.22	-0.44
MPY2	-0.13	0.07	-0.4	-0.08	-0.01	-0.13	0.32	0.22	0.44	-0.18	-0.32	0.09	-0.24	0.18
MPY3	0.06	-0.27	-0.2	-0.02	0.02	-0.12	0.23	0.33	0.11	0	0.05	0.12	-0.17	-0.11
VMO1	0.27	0.35	0.08	0.28	0.28	-0.48	-0.34	-0.09	0.08	0.39	0.48	0.18	-0.14	0.02
VMO2	0.38	0.34	0.26	0.44	0.44	-0.33	-0.14	0.09	0.2	0.29	0.3	0.12	0	-0.1
VMO3	0.36	0.30	0.02	0.24	0.22	-0.4	-0.26	0.06	0.24	0.27	0.13	0.43	-0.04	-0.04
VMO4	0.36	0.41	-0.34	-0.12	-0.15	-0.74	-0.4	-0.39	0.01	0.43	-0.01	0.58	-0.25	-0.11
VKÄ1	0.2	0.46	-0.04	0.06	0.04	-0.32	-0.2	-0.03	-0.03	0.33	0.22	0.42	0.15	0.08
VKÄ2	0.51	0.44	0.21	0.12	0.05	-0.25	-0.22	-0.36	-0.07	0.48	0.32	0.04	0.04	-0.02

	MKU1	MKU2	MKU3	MKU4	SOS1	SOS2	SOS3	MPY1	MPY2	MPY3	VMO1	VMO2	VMO3	VMO4	VKÄ1	VKÄ2
Ikä	-0.48	-0.43	0.04	0.13	-0.48	0.13	0.09	0.52	-0.13	0.06	0.27	0.38	0.36	0.36	0.2	0.51
Kokemus	-0.44	-0.41	-0.45	-0.37	-0.14	-0.02	0.18	0.39	0.07	-0.27	0.35	0.34	0.3	0.41	0.46	0.44
MAL1	0.18	0.05	0.2	0.45	-0.19	0.01	0.05	0	-0.4	-0.20	0.08	0.26	0.02	-0.34	-0.04	0.21
MAL2	-0.06	-0.14	-0.03	0.08	-0.01	-0.1	0.12	0.13	-0.08	-0.02	0.28	0.44	0.24	-0.12	0.06	0.12
MAL3	-0.04	-0.12	-0.06	0.07	0.02	-0.12	0.08	0.14	-0.01	0.02	0.28	0.44	0.22	-0.15	0.04	0.05
MVA1	0.47	0.23	0.26	0.49	-0.16	-0.31	-0.32	-0.32	-0.13	-0.12	-0.48	-0.33	-0.4	-0.74	-0.32	-0.25
MVA2	0.2	0.02	0.14	0.41	-0.37	-0.33	-0.36	0.11	0.32	0.23	-0.34	-0.14	-0.26	-0.4	-0.2	-0.22
MVA3	-0.14	-0.18	0.12	0.11	-0.02	-0.25	-0.14	0.12	0.22	0.33	-0.09	0.09	0.06	-0.39	-0.03	-0.36
MUH1	-0.14	-0.1	0.07	0.01	-0.21	-0.28	-0.07	0.28	0.44	0.11	0.08	0.2	0.24	0.01	-0.03	-0.07
MUH2	-0.09	0.02	0.3	0.25	-0.05	0.39	0.25	0.22	-0.18	0.00	0.39	0.29	0.27	0.43	0.33	0.48
MUH3	-0.09	-0.05	0.23	0.07	0.25	0.63	0.12	0.02	-0.32	0.05	0.48	0.3	0.13	-0.01	0.22	0.32
MTE1	-0.42	-0.19	0.01	-0.24	0.06	-0.03	0.53	0.23	0.09	0.12	0.18	0.12	0.43	0.58	0.42	0.04
MTE2	0.25	0.12	0.27	0.42	0.06	-0.14	-0.11	-0.22	-0.24	-0.17	-0.14	0	-0.04	-0.25	0.15	0.04
MTE3	0.28	0.35	0.24	0.1	0.29	-0.08	-0.42	-0.44	0.18	-0.11	0.02	-0.1	-0.04	-0.11	0.08	-0.02
MKU1	1	0.84	0.45	0.61	0.06	-0.21	-0.32	-0.65	-0.16	-0.42	-0.59	-0.60	-0.67	-0.65	-0.59	-0.27
MKU2	0.84	1	0.62	0.50	0.13	-0.34	-0.31	-0.69	0.09	-0.32	-0.62	-0.70	-0.59	-0.46	-0.69	-0.39
MKU3	0.45	0.62	1	0.75	-0.11	0.05	-0.16	-0.29	-0.09	-0.04	-0.39	-0.45	-0.33	-0.24	-0.49	0.01
MKU4	0.61	0.50	0.75	1	-0.38	-0.02	-0.08	-0.15	-0.4	-0.21	-0.48	-0.35	-0.48	-0.43	-0.41	0.13
SOS1	0.06	0.13	-0.11	-0.38	1	0.02	0.05	-0.3	0.15	0.21	0.35	0.23	0.22	0.12	0.18	-0.34
SOS2	-0.21	-0.34	0.05	-0.02	0.02	1	0.1	0.2	-0.52	0.01	0.64	0.42	0.13	0.2	0.3	0.75
SOS3	-0.32	-0.31	-0.16	-0.08	0.05	0.1	1	0.45	-0.41	-0.02	0.44	0.59	0.4	0.33	0.34	0.24
MPY1	-0.65	-0.69	-0.29	-0.15	-0.30	0.2	0.45	1	0	0.18	0.50	0.67	0.54	0.56	0.41	0.38
MPY2	-0.16	0.09	-0.09	-0.4	0.15	-0.52	-0.41	0	1	0.48	-0.23	-0.28	-0.03	0.12	-0.27	-0.61
MPY3	-0.42	-0.32	-0.04	-0.21	0.21	0.01	-0.02	0.18	0.48	1	0.14	0.14	0.01	0.13	-0.07	-0.29
VMO1	-0.59	-0.62	-0.39	-0.48	0.35	0.64	0.44	0.50	-0.23	0.14	1	0.91	0.68	0.57	0.69	0.53
VMO2	-0.60	-0.70	-0.45	-0.35	0.23	0.42	0.59	0.67	-0.28	0.14	0.91	1	0.72	0.50	0.70	0.43
VMO3	-0.67	-0.59	-0.33	-0.48	0.22	0.13	0.4	0.54	-0.03	0.01	0.68	0.72	1	0.74	0.80	0.2
VMO4	-0.65	-0.46	-0.24	-0.43	0.12	0.2	0.33	0.56	0.12	0.13	0.57	0.50	0.74	1	0.63	0.37
VKÄ1	-0.59	-0.69	-0.49	-0.41	0.18	0.3	0.34	0.41	-0.27	-0.07	0.69	0.70	0.80	0.63	1	0.43
VKÄ2	-0.27	-0.39	0.01	0.13	-0.34	0.75	0.24	0.38	-0.61	-0.29	0.53	0.43	0.2	0.37	0.43	1