Helmi Pynnönen

TEKOÄLYN ROOLI DISINFORMAATION LEVIÄMISESSÄ JA TORJUMISESSA SOSIAALISESSA MEDIASSA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2024

TIIVISTELMÄ

Pynnönen, Helmi
Tekoälyn rooli disinformaation leviämisessä ja torjumisessa sosiaalisessa mediassa
Jyväskylä: Jyväskylän yliopisto, 2024, 34 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja(t): Kokko, Tuomas

Tekoälyn kehitys on tehnyt disinformaation eli harhaanjohtavan tiedon tahallisesta levittämisestä tehokasta ja vaikeasti tunnistettavaa. Erilaiset tekoälyteknologiat, kuten koneoppiminen ja syväoppiminen, mahdollistavat sisällöntuotannon manipuloinnin sekä disinformaation tehokkaan levittämisen sosiaalisen median alustoilla. Lisäksi sosiaalinen media tarjoaa otollisen ympäristön disinformaation levittämiselle tiedon faktantarkistuksen ollessa alustoilla haastavaa. Toisaalta tekoälyteknologioita voidaan kuitenkin hyödyntää myös disinformaation tunnistamiseen ja torjumiseen sosiaalisessa mediassa. Tämän kirjallisuuskatsauksen tavoitteena oli ymmärtää tekoälyn rooli sekä disinformaation leviämisessä että sen torjumisessa sosiaalisen median kontekstissa. Tutkimuksessa todettiin tekoälypohjaisten syvävääreennösten, sosiaalisten bottien ja mikrokohdentamisen olevan tyypillisiä disinformaation levittämiseen hyödynnettyjä teknologioita. Toisaalta tutkimustulosten perusteella erilaiset kone- ja syväoppimismenetelmät mahdollistavat myös disinformaation torjumisen sosiaalisessa mediassa. Tämä tutkimus toteutettiin kuvailevana kirjallisuuskatsauksena, joka perustuu 42 lähteen analyysiin.

Asiasanat: tekoäly, disinformaatio, sosiaalinen media, sosiaaliset botit, syvävääreennökset, mikrokohdentaminen

ABSTRACT

Pynnönen, Helmi
The role of artificial intelligence in the spread and prevention of disinformation on social media
Jyväskylä: University of Jyväskylä, 2024, 34 pp.
Information Systems, Bachelor's Thesis
Supervisor(s): Kokko, Tuomas

The development of artificial intelligence has made the spread of disinformation, defined as the intentional distribution of misleading information effective and difficult to detect. Different artificial intelligence technologies, such as machine learning and deep learning, enable the manipulation of content production and the effectively spread disinformation on social media platforms. Additionally, social media provides a suitable environment for disseminating disinformation, while fact-checking information on the platforms is challenging. Conversely, artificial intelligence technologies can also be employed to identify and combat disinformation on social media. This literature review aims to understand the role of artificial intelligence in both the spread of disinformation and the efforts to combat it in the context of social media. The study found that AI-powered deepfakes, social bots, and microtargeting are typical technologies used to spread disinformation. However, based on the research results, various machine learning and deep learning methods can also enable the fight against disinformation on social media. This research was completed as a descriptive literature review based on the analysis of 42 sources.

Keywords: artificial intelligence, disinformation, social media, social bots, deepfakes, microtargeting

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Tekoälyteknologioiden kehitys on muuttanut viestinnän keinoja sosiaalisessa mediassa mahdollistaen esimerkiksi sisällöntuotannon ja sisällönpersonoinnin paremman käyttäjäkokemuksen mahdollistamiseksi (Shankar, 2024). Yhä enenevissä määrin keskustelua aiheuttaa kuitenkin myös tekoälyn käyttö disinformaation levittämisen mahdollistajana.

Disinformaatiolla viitataan tahallisesti levitettyyn harhaanjohtavaan tietoon, jonka tarkoituksena on aiheuttaa yleistä vahinkoa (Fallis, 2015). Disinformaatio voi johtaa esimerkiksi luottamuksen puutteeseen mediaa tai instituutioita kohtaan ja edistää yhteiskunnan jakautumista (Kertysova, 2018). Verkossa esiintyvä disinformaatio ei ole uusi ilmiö, mutta tekoälyn kehittyminen on tuonut uusia mahdollisuuksia sen levittämiseksi internetissä (Kertysova, 2018).

Tekoälyteknologiat voivatkin altistaa manipuloidulle tiedolle, puolueellisille algoritmeille sekä tiukalle sisällön personoinnille edistäen disinformaation ongelman lisääntymistä verkossa (Kertysova, 2018). Tyypillisiä verkossa esiintyviä tekoälypohjaisia disinformaation levittämiseen hyödynnettyjä tekniikoita ovat syvävääreennökset, sosiaaliset botit sekä mikrokohdentaminen (Bontridder & Poullet, 2021). Disinformaation levittäminen syvävääreennettyjen kuvien ja videoiden välityksellä on sosiaalisessa mediassa tyypillistä (Shu ym., 2020). Lisäksi tekoälyyn pohjautuvia sosiaalisia botteja (Hajli ym., 2022) sekä sisällön personointiin hyödynnettyä mikrokohdentamista voidaan käyttää disinformaation levittämisen tehostamiseen sosiaalisessa mediassa (Bontridder & Poullet, 2021). Tekoälyteknologiat ovat nykyisin myös yhä tiiviimmin osa arkipäivää. Esimerkiksi syvävääreennösteknologiat ovat nykyisin yhä helpommin saavutettavissa (Kertysova, 2018; Shu ym., 2020), mikä voi luoda uusia mahdollisuuksia disinformaation levittämiselle myös yksilötasolla.

Lisäksi manuaalinen faktantarkistus on sosiaalisessa mediassa usein tehotonta suuresta datan määrästä johtuen (Rastogi & Bansal, 2022). Tämä luo tarvetta teknologiselle ja automatisoidulle faktantarkistukselle. Erilaiset tekoälyteknologiat voivatkin puolestaan tarjota myös mahdollisuuksia disinformaation tunnistamiseen ja torjumiseen, edistäen sosiaalisen median palveluiden turvallisuutta (Shankar, 2024).

Tämän tutkielman tavoitteena on tarkastella tekoälyn kaksinaista roolia disinformaation levittäjänä ja torjujana sosiaalisen median kontekstissa. Tutkielma vastaa tutkimuskysymykseen: ”Miten tekoäly vaikuttaa disinformaation leviämiseen sosiaalisessa mediassa, ja millaisia mahdollisuuksia se voi tarjota tämän ehkäisemiseksi?”. Ilmiö on suhteellisen uusi, minkä vuoksi siitä on rajallisesti aiempaa tutkimusta. Ilmiön ymmärtäminen on tärkeää niin tekoälyteknologioiden vastuullisen kehittämisen, disinformaation torjunnan, sosiaalisen median palveluiden turvallisuuden parantamisen kuin myös poliittisen päätöksenteon tukemisen näkökulmasta. Lisäksi tutkimus voi lisätä digi- ja medialukutaitoa ja siten ehkäistä disinformaation uhkaa sosiaalisessa mediassa.

Tutkimus on toteutettu kuvailevana kirjallisuuskatsauksena. Tutkimuksessa kootaan aiempaa kirjallisuutta pyrkimyksenä antaa kokonaisvaltainen ymmärrys tutkittavasta ilmiöstä. Synteesiä tässä tutkimuksessa muodostuu aiemman kirjallisuuden määritelmiä ja johtopäätelmiä yhdistelemällä.

Tutkielman lähdeaineisto on kerätty Jykdok-, Google Scholar-, IEEE Xplore- ja Scopus-tietokannoista. Aineiston etsimisessä on hyödynnetty myös Keenious-työkalua. Lähdeaineiston luotettavuutta ja laatua on pyritty analysoimaan tarkastelemalla Julkaisufoorumin laatuluokituksia. Tutkielman aineisto koostuu 42 lähteestä, joista 40 on julkaistu Julkaisufoorumissa laatuluokituksella 1–3. Tekoälyohjelmaa ChatGPT on hyödynnetty tutkielman aineiston sisällöllisessä analysoinnissa. Kaikki tutkielmassa käytetyt lähteet on kuitenkin luettu ja analysoitu itse.

Julkaisufoorumi ei ole luokitellut Neittaanmäen ym. (2019) kirjaa ”Tekoälyn perusteita ja sovelluksia” tai McCarthyn (2004) artikkelia ”What is artificial intelligence?”. Neittaanmäen ym. (2019) Jyväskylän yliopistossa julkaistua kirjaa on hyödynnetty tutkielmassa kokonaisvaltaisen käsityksen muodostamiseksi tekoälyn ilmiöstä. McCarthyn (2004) artikkelin merkittävyydestä taas kertoo suuri viittauksien määrä.

Myös viittausten määrää on siis pyritty tarkastelemaan lähteiden merkittävyyttä arvioidessa. Tutkittavasta ilmiöstä on kuitenkin rajallisesti tutkimusaineistoa, minkä vuoksi viittausten määrää on harvoin pystytty pitämään lähteen valinnan ratkaisevana kriteerinä.

Tutkielmassa käytetyt lähteet on pääosin etsitty seuraavilla hakusanoilla: ”artificial intelligence”, ”disinformation” ja ”social media”, sekä näiden sanojen yhdistelmillä. Lisäksi hakusanoja ”social bots”, ”deepfakes”, ja ”microtargeting” on hyödynnetty myöhemmin hakuprosessissa. Tutkielman aiheeseen liittyvää tutkimusta on pääosin tehty englanniksi, minkä vuoksi myös hakusanat ovat englanninkielisiä.

Tutkielma sisältää kuusi lukua; johdannon, neljä sisältölukua ja yhteenvedon. Ensimmäisessä luvussa eli johdannossa esitellään tutkimusaihe, tutkimuskysymys sekä tutkimuksen toteutustapa. Tutkielman toisessa luvussa eli ensimmäisessä sisältöluvussa määritellään disinformaation ja sosiaalisen median käsitteet sekä tuodaan esille käsitteisiin liittyviä olennaisia piirteitä. Lisäksi luvussa käsitellään disinformaation ilmenemistä sosiaalisessa mediassa yleisluontoisesti. Kolmannessa luvussa eli toisessa sisältöluvussa määritellään tekoälyn,

koneoppimisen sekä syväoppimisen käsitteet. Neljännessä luvussa, eli kolmannessa sisältöluvussa tarkastellaan tarkemmin tekoälyn roolia disinformaation levittämisen näkökulmasta sosiaalisen median kontekstissa. Viidennessä luvussa, eli viimeisessä sisältöluvussa taas käsitellään tekoälyn roolia disinformaation torjumisen näkökulmasta sosiaalisessa mediassa. Lisäksi luvussa sivutaan tekoälyteknologioiden käyttöön liittyviä keskeisiä haasteita disinformaation torjunnan yhteydessä. Viimeisessä luvussa kootaan vielä yhteen tutkimuksen pääpiirteet ja tulokset sekä tuodaan esille mahdollisia jatkotutkimusaiheita. Lisäksi luvussa arvioidaan tutkimuksen rajoitteita ja validiteettia.

2 DISINFORMAATIO SOSIAALISESSA MEDIASSA

Tässä luvussa käsitellään disinformaatiota ja sen ilmenemistä sosiaalisessa mediassa. Luvun ensimmäisessä alaluvussa tarkastellaan disinformaation määritelmää sekä pyritään ymmärtämään disinformaation ja misinformaation ero. Toisessa alaluvussa määritellään sosiaalisen media sekä käsitellään disinformaation ilmenemistä sosiaalisen mediassa.

2.1 Disinformaatio käsitteenä

Disinformaatiolla tarkoitetaan tahallisesti levitettyä harhaanjohtavaa, manipuloitua tai väärennettyä tietoa, jolla pyritään aiheuttamaan yleistä vahinkoa tai tavoittelemaan taloudellista voittoa (Fallis, 2015; Saurwein & Spencer-Smith, 2020). Fetzer (2004) määrittelee disinformaation tahallisena väärän tai harhaanjohtavan tiedon levittämisenä, jonka tarkoituksena on johtaa harhaan, hämmentää tai pettää. Disinformaatio voi olla joko tahallista tunnetusti harhaanjohtavan, väärän tiedon levittämistä tai tahallista oikean tiedon vääristämistä (Fetzer, 2004). Disinformaation tyypillisiä esiintymismuotoja on muun muassa väärennetyt kuvat ja asiakirjat, harhaanjohtava mainonta, väärennetty verkkosisältö sekä valtiollinen propaganda (Fallis, 2015). Propaganda voidaan jakaa valkoiseen, mustaan ja harmaaseen propagandaan (Becker, 1949; Guth, 2009). Musta propaganda perustuu valheelliseen informaatioon, jonka tarkoituksena on johtaa harhaan (Guth, 2009). Valkoinen propaganda taas perustuu tunnistettuihin lähteisiin, ja harmaassa propagandassa informaation lähteeseen ja sen totuudenmukaisuuteen liittyy epäselvyyksiä (Guth, 2009). Tämän jaottelun perusteella musta propaganda liittyy selkeinten disinformaation määritelmään, sillä molempien tarkoituksena on johtaa harhaan valheellisella informaatiolla.

Fallisin (2015) mukaan disinformaatio on hyvin tyypillisesti taloudellista, poliittista tai lääketieteellistä. Tällainen disinformaatio voi johtaa fyysisiin, emotionaalisiin ja taloudellisiin vahinkoihin (Fallis, 2015). Bradshaw ja DeNardis (2024) lisäävät Fallisin määritelmään digitaalisen infrastruktuuriin kohdistuvan

disinformaation, jolla viitataan kriittisen internet-infrastruktuurin tahalliseen manipulointiin. Tällä voidaan tarkoittaa esimerkiksi verkkotunnusjärjestelmän (DNS) tai julkisen avaimen infrastruktuurin (PKI) manipulointia (Bradshaw & DeNardis, 2024).

Disinformaatio voi heikentää luottamusta tiedotusvälineisiin, mediaan ja instituutioihin, vaikuttaa äänestyspäätöksiin ja poliittisiin mielipiteisiin sekä syventää yhteiskunnan jakautumista (Kertysova, 2018). Euroopan Unionin toimielimet pitävätkin disinformaatiota laillisin keinoin ehkäistävänä ilmiönä sen haitallisuuden vuoksi (Bontridder & Poullet, 2021).

Ahmadin ym. (2021) mukaan disinformaatio muodostuu kolmesta tekijästä. Näitä tekijöitä ovat luodun ja jaetun tiedon muoto (esimerkiksi keksitty informaatio tai väärä konteksti), sisällöntuottajan motiivi ja saatu hyöty (esimerkiksi taloudellinen tai poliittinen) sekä sisällön levitystapa (esimerkiksi sosiaalinen media tai uutiset) (Ahmad ym., 2021). Fetzer taas jakaa disinformaation viidelle eri tasolle (2004). Näitä tasoja ovat hänen mukaansa tiedon esittäjän epäpätevyys käsitellä aihetta, merkityksellisen todisteen tahallinen huomiotta jättäminen, hyökkäys teoksen tekijää vastaan harhaanjohtavin perustein, teoksen esittäminen virheellisesti jättäen tahallisesti huomiotta sen tärkeimpiä ominaisuuksia sekä tarkoituksellisesti harhaanjohtavan tiedon esittäminen jättämällä tietyt tiedot huomiotta esimerkiksi nostamalla esille ainoastaan asiaa puoltavia argumentteja (Fetzer, 2004).

Kaikissa edellä mainituissa disinformaation määritelmissä korostuu tahallinen harhaanjohtaminen. Disinformaatio voidaan siis yksinkertaistaen määritellä informaatioksi, joka on tarkoituksenmukaisesti harhaanjohtavaa ja luotu vahingoittamaan (Fallis, 2015; Saurwein & Spencer-Smith, 2020). Nämä ovat piirteitä, jotka erottavat disinformaation misinformaatiosta.

Misinformaatio on väärän tiedon levittämistä ilman tarkoitusta johtaa harhaan tai vahingoittaa (Aïmeur ym. 2023; Fallis, 2015; Kertysova, 2018; Saurwein & Spencer-Smith, 2020; Shu ym., 2020). Kaikkia harhaanjohtavia väitteitä ei siis voi pitää disinformaationa, vaan harhaanjohtava tieto voi olla vilpittömästi tahatonta tai johtua tietämättömyydestä (Ahmad ym., 2021). Disinformaation levittäjä onkin tietoinen tiedon harhaanjohtavuudesta (Fallis, 2015), kun taas misinformaation levittäminen johtuu usein rehellisestä vahingosta (Shu ym., 2020). Vaikka misinformaatiokin voi johtaa harhaan, disinformaation aiheuttama vahinko on yleensä huomattavasti vakavampaa (Fallis, 2015).

Disinformaation ja misinformaation erottaminen voi olla vaikeaa, sillä tiedon levittämisen takana olevan agendan määrittely voi olla haasteellista (Ahmad ym., 2021). Lisäksi disinformaation erottaminen misinformaatiosta on haastavaa tapauksissa, joissa informaatiolla pyritään ylläpitämään tietämättömyyttä, disinformaatio johtaa harhaan väärää kohdeyleisöä tai johtaa harhaan eri tavoin kuin on ollut pyrkimyksenä (Fallis, 2015). Disinformaatiota määritellessä tulee lisäksi ottaa huomioon, että se ei välttämättä mustavalkoisesti johda virheellisiin uskomuksiin, sillä tiedon vastaanottaja voi myös olla uskomatta tietoa (Fallis, 2015). Taulukossa 1 havainnoidaan tässä luvussa esiteltyjä disinformaation ja misinformaation tyypillisiä eroja.

TAULUKKO 1 Disinformaation ja misinformaation tyypilliset erot

Ominaisuus	Disinformaatio	Misinformaatio
Tieto on harhaanjohtavaa	✓	✓
Tiedon levittäminen on tahallista	✓	✗
Tiedon levittäjä on tietoinen tiedon harhaanjohtavuudesta	✓	✗
Levittämistä ohjaa jokin päämäärä	✓	✗
Pyrkimyksenä vahingon aiheuttaminen	✓	✗
Voi aiheuttaa laajaa vahinkoa	✓	✗

Disinformaation määrittely on tärkeää, jotta voidaan ymmärtää sen leviämistapoja sekä tunnistamisen ja torjumisen keinoja (Fallis, 2015). Lisäksi disinformaation uhka on kasvanut teknologisen kehityksen myötä teknologioiden helpottaessa disinformaation leviämistä ja luomista (Fallis, 2015). Teknologiat ovat siis tuoneet uusia haasteita disinformaation leviämiselle, ja tämän vuoksi ilmiön tutkimista voidaan edelleen pitää ajankohtaisena.

2.2 Disinformaatio sosiaalisessa mediassa

Disinformaation luominen ja levittäminen sosiaalisessa mediassa on yleistyvä ongelma (Shu ym., 2020). Sosiaalisella medialla tarkoitetaan internet-pohjaisia kanavia, joilla käyttäjät voivat kommunikoida toistensa kanssa asynkronisesti tai reaaliajassa (Carr & Hayes, 2015). Sosiaalisen median palveluiden keskiössä on käyttäjien luoma sisältö sekä vuorovaikutus käyttäjien välillä (Carr & Hayes, 2015; Obar & Wildman, 2015). Vuorovaikutus sosiaalisen median alustoilla toteutuu yleensä tykkäysten, sisällön jakamisen ja kommentoinnin välityksellä (Obar & Wildman, 2015). Lisäksi palvelut ovat tavallisesti ilmaisia ja helppokäyttöisiä (Ahmad ym., 2021). Tämän seurauksena sosiaalisen median palvelut ovat helposti saatavilla suurelle käyttäjäkunnalle. Sosiaalisen median alustat on tyypillisesti suunniteltu läheisten kanssa kommunikointiin, mutta niitä käytetään nykyisin myös omien mielipiteiden sekä harhaanjohtavan tiedon jakamiseen (Ahmad ym., 2021).

Sosiaalinen media mahdollistaakin nopean informaation jakamisen, oli tieto oikeellista tai harhaanjohtavaa (Shu ym., 2020). Sosiaalisen median käyttäjät

voivatkin osallistua niin disinformaation kuluttamiseen, kuin myös sen levittämiseen (Saurwein & Spencer-Smith, 2020). Sosiaalisen median alustoilla disinformaatio leviää tyypillisesti kuvien, videoiden ja tekstin muodossa (Shu ym., 2020). Sosiaalisen median alustoja on esimerkiksi Facebook, Twitter (nykyinen X) ja Instagram (Orabi ym., 2020).

Shu ym. (2020) painottaa sosiaalisessa mediassa disinformaatioon ja sen tunnistamiseen liittyvän niin sisällöllisiä kuin käyttäjiin liittyviä haasteita. Heidän mukaansa disinformaatiolle tyypillistä on huomiota herättävien sekä tunnelatautuneiden ilmaisujen hyödyntäminen, pyrkimyksenä saada sosiaalisen median käyttäjät reagoimaan sisältöön. Lisäksi tunneperäiset syyt, kuten aiheen henkilökohtaisuus tai ahdistavuus tekee sosiaalisen median käyttäjistä haavoittuvampia harhaanjohtavan informaation hyväksymiselle ja levittämiselle (Shu ym., 2020). Myös tietoisuus disinformaatiosta on vähäistä, ja on tyypillistä, että ihmiset luottavat trendikkäältä vaikuttavaan tietoon (Shu ym., 2020).

Saurwein ja Spencer-Smithin (2020) mukaan käyttäjiin ja sisältöön liittyvien haasteiden lisäksi sosiaalisen median palveluiden rakenne tukee disinformaation leviämistä suositusalgoritmien, tykkäys- ja jakamispainikkeiden sekä automaattisen sisällön toiston muodossa. Näin palveluiden rakenne mahdollistaa käyttäjien vuorovaikutuksen disinformaatiota sisältävän sisällön kanssa. Tämä vuorovaikutus taas edistää disinformaation leviämistä algoritmien suosiessa suosittua sisältöä (Saurwein & Spencer-Smith, 2020). Sosiaalisessa mediassa erityisenä disinformaation esiintymisen haasteena on myös suositusjärjestelmistä johtuvien suodatinkuplien (engl. filter bubbles) sekä kaikukammioiden (engl. echo chambers) muodostuminen (Shu ym., 2020). Tällöin sosiaalisen median käyttäjille tarjotaan heidän näkemyksiään vastaavaa sisältöä, mikä vahvistaa tietynlaista ajattelutapaa ottamatta huomioon näkemysten vastakkaista puolta (Shu ym., 2020).

Toisaalta sosiaalisen median yrityksillä on myös mahdollisuus vaikuttaa alustojen sääntelyyn, ja näin pyrkiä ehkäisemään disinformaatiota (Saurwein & Spencer-Smith, 2020). Faktojen tarkistaminen sosiaalisessa mediassa on kuitenkin vaikeaa, sillä kuka tahansa voi jakaa ja luoda sisältöä alustoilla (Ahmad ym., 2021; Saurwein & Spencer-Smith, 2020). Lisäksi suuri datan määrä luo haasteita manuaaliselle faktantarkistukselle sosiaalisessa mediassa (Rastogi & Bansal, 2022).

Disinformaation torjuminen ei välttämättä myöskään näyttäydy sosiaalisen median alustoille taloudellisesti houkuttelevana. Disinformaatiota sisältävät postaukset herättävät huomiota ja lisäävät käyttäjien sitoutumista tukien alustojen liiketoimintamallia (Walker ym., 2019). Myös tekoälyyn pohjautuvat suositusalgoritmit mahdollistavat sisällön personoinnin lisäten käyttäjien tyytyväisyyttä ja sitoutuneisuutta palvelun käyttöön (Shankar, 2024). Lisäksi sosiaalisen median suositusjärjestelmiä käytetään myös mainontaan (Shankar, 2024).

3 TEKOÄLY

Tässä luvussa tarkastellaan tekoälyn määritelmää. Tämän lisäksi luvussa tarkastellaan tekoälyn osa-alueita ja näiden ominaisuuksia tarkemmin. Ensimmäisessä alaluvussa käsitellään tekoälyn määritelmää yleisesti. Toisessa alaluvussa käsitellään koneoppimista ja koneoppimisen keskeisiä piirteitä. Kolmannessa alaluvussa käsitellään vielä syväoppimista ja sen yhteyttä keinotekoisiin syviin neuroverkkoihin. Koneoppimisen ja syväoppimisen perusteiden käsittely on tutkielman aiheen kannalta tärkeää, sillä suuri osa tässä tutkielmassa käsiteltävistä disinformaation leviämisen ja torjumisen keinoista pohjautuu näihin teknologioihin.

3.1 Tekoäly käsitteenä

Wangin (2019) määritelmän mukaan tekoälyllä tarkoitetaan järjestelmää, joka mukailee ihmisälyn kaltaisia kognitiivisia toimintoja, kuten esimerkiksi oppimista, ajattelua sekä ongelmanratkaisua. Tästä määritelmästä huolimatta tekoäly voidaan määritellään hieman eri tavoin asiayhteydestä riippuen, eikä sille ole yhtä oikeaa määritelmää (Abbass, 2021; Neittaanmäki ym., 2019; Wang, 2019). Selkeä termin määrittely voi kuitenkin selkeyttää tutkimusta asettaen tutkimukselle rajoja ja tavoitteita (Abbass, 2021; Wang, 2019).

Eräs tapa pyrkiä tutkimaan tekoälyn määritelmää on tarkastella sanoja teko (keinotekoinen) ja älykkyys erillisinä termeinä (Abbass, 2021). Älykkyyden määritelmän vakiintumattomuus on tehnyt tekoälyn määrittelystä haastavaa (Wang, 2019). Tämän lisäksi koneen älykkyyden ajatellaan eroavan ihmisen älykkyydestä (McCarthy, 2004; Shevlin ym., 2019; Wang, 2019). Yksinkertainen tapa määritellä älykkyys on nähdä se kykynä pystyä saavuttamaan tavoitteita muuttuvissa ympäristöissä (Shevlin ym., 2019). Sanalla teko taas tyypillisesti viitataan keinotekoiseen, ihmisen tuottamaan tai luontoa imitoivaan (Abbass, 2021).

Haenleinin ja Kaplanin määritelmän mukaan tekoäly voidaan käsittää ”järjestelmän kykynä tulkita ulkoista dataa, oppia tällaisesta datasta sekä käyttää tätä tietoa tehtävien sekä tavoitteiden saavuttamiseksi joustavan mukautumisen avulla” (Haenlein & Kaplan, 2019, s. 1). Myös McCarthy (2004) korostaa tekoälyn määritelmässään tietokoneiden käyttämistä ihmisen älykkyyden ymmärtämiseksi, tarvitsematta kuitenkaan rajoittua biologisesti luonnossa havaittavissa oleviin menetelmiin. Älykkäät järjestelmät eroavatkin perinteisestä laskennasta (engl. traditional computing) siten, että ne pystyvät dynaamiseen ongelmanratkaisuun sopeutuen muuttuviin olosuhteisiin sekä tekemään johtopäätöksiä epätäydellisen informaation perusteella (Wang, 2019).

Neittaanmäen ym. (2019) mukaan tekoäly voidaan lisäksi jakaa heikkoon ja vahvaan tekoälyyn. Heikolla tekoälyllä tarkoitetaan yksinkertaisista tehtävistä suoriutuvia algoritmeja. Vahvalla tekoälyllä viitataan taas tekoälyyn, joka toimii irrallaan ihmisälystä ja on kykenevä ennustamaan tulevaisuuteen. (Neittaanmäki ym., 2019.) Toinen tapa jakaa tekoäly on jakaa se logiikkavaikutteiseen- sekä aivovaikutteiseen lähestymistapaan (LeCun ym., 2021). Logiikkavaikutteisessa paradigmassa älykkyyden keskiössä on sääntöpohjainen päättely, kun taas aivovaikutteisessa datasta oppiminen (LeCun ym., 2021).

Tekoälyä hyödynnetään nykyisin monissa arkipäiväisissä teknologioissa, kuten hakukoneissa, kohdennetussa mainonnassa ja kameroiden kasvojentunnistuksessa (Neittaanmäki ym., 2019). Haenlein ja Kaplan (2019) uskovatkin, että tekoäly tulee vaikuttamaan merkittävästi ihmisten elämään niin yksilötasolla kuin organisaatioissa verraten ilmiön laajuutta internettiin ja sosiaaliseen mediaan

3.2 Koneoppiminen

Koneoppiminen (engl. machine learning) tarkoittaa koneen dynaamista kykyä oppia sille toimitetusta datasta ja algoritmeista, sekä kykyä tehdä päätelmiä tämän tiedon perusteella muokaten itseään ilman nimenomaista ohjelmointia (Jakhar & Kaur, 2020). Koneoppimisen algoritmit pyrkivät minimoimaan virheitä pyrkien kuitenkin aina mahdollisimman todennäköiseen lopputulemaan (Jakhar & Kaur, 2020). Koneoppimisen tavoitteena on usein tiedon tulkinnan automatisointi sekä koneen havainnointikyvyn laajentaminen (Neittaanmäki ym., 2019). Koneoppimisessa algoritmien valinta riippuu muun muassa oppimistyylistä, datan määrästä, tarvittavasti tallennustilasta sekä ratkaisun tehokkuuden vaatimuksista (Neittaanmäki ym., 2019).

Koneoppiminen voidaankin tyypillisesti jakaa kolmeen eri luokkaan oppimisen tyylin eli oppimisparadigman perusteella (Jordan & Mitchell, 2015; Neittaanmäki ym., 2019). Nämä luokat ovat ohjattu oppiminen (engl. supervised learning), ohjaamaton oppiminen (engl. unsupervised learning) sekä vahvistusoppiminen (engl. reinforcement learning) (Jordan & Mitchell, 2015; Neittaanmäki ym., 2019).

Ohjatussa oppimisessa hyödynnetään merkittyä dataa (engl. labeled data) algoritmien kouluttamiseen lopputulemien ennustamiseksi ja tunnistamiseksi syöte-tavoite-pareja hyödyntämällä (Jordan & Mitchell, 2015). Neittaanmäen ym. (2019) mukaan ohjattu oppiminen perustuu syöte-tavoite-pareista koostuvaan aineistoon tavoitteena koneen kyky tehdä jaottelu samankaltaisille aineistoille. Lisäksi ohjattu oppiminen voidaan Neittaanmäen ym. mukaan jakaa datan tyypin perusteella joko luokitteluksi tai regressioksi. Luokittelusta puhutaan silloin, kun data on diskreettiä eli data voidaan jaotella ryhmiin. Regressioon viitataan taas silloin, kun data on jatkuvaa. (Neittaanmäki ym., 2019.)

Ohjaamattomassa oppimisessa datasta pyritään tunnistamaan syötteiden välisiä riippuvuuksia ja samankaltaisuuksia (Neittaanmäki ym., 2019). Ohjaamaton oppiminen eroaa ohjatusta oppimisesta siten, että ohjaamaton oppiminen perustuu ei-merkittyyn dataan (engl. unlabeled data) (Jordan & Mitchell, 2015). Tällöin data klusteroidaan eli ryhmitellään datan ominaisuuksien perusteella erilaisiin samankaltaisten syötteiden ryhmiin (Neittaanmäki ym., 2019).

Vahvistusoppimisessa oppiminen perustuu taas ympäristön kanssa vuorovaikutuksesta saataviin niin sanottuihin palkkiosignaaleihin (Jordan & Mitchell, 2015). Vahvistusoppimista määrittelee siis koneen ympäristöstä saadun positiivisen tai negatiivisen palautteen määrä (Jordan & Mitchell, 2015; Neittaanmäki ym., 2019). Tällöin algoritmi pyrkii löytämään ratkaisun, josta ympäristö antaa mahdollisimman positiivista palautetta (Jordan & Mitchell, 2015; Neittaanmäki ym., 2019). Vahvistusoppimisessa yhdistyy ohjattu ja ohjaamaton oppiminen (Jordan & Mitchell, 2015). Tuleekin ottaa huomioon, että nykyinen tutkimus yhdistelee usein edellä mainittuja kolmea oppimisparadigmaa sekaisin (Jordan & Mitchell, 2015).

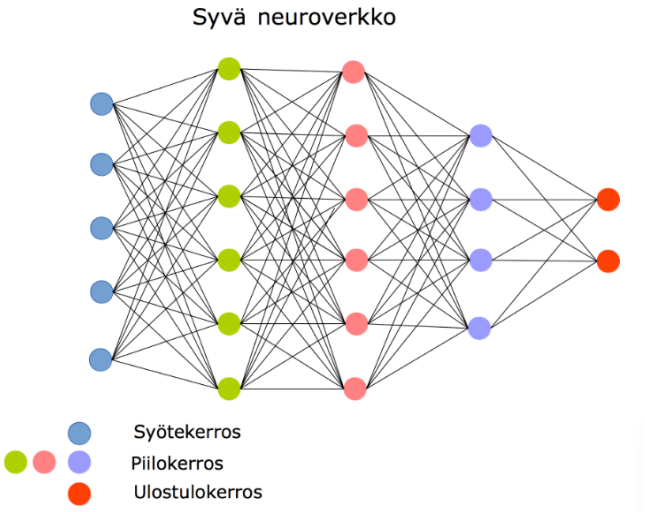
3.3 Syväoppiminen

Syväoppiminen mahdollistaa tehokkaan suurten datamäärien prosessoinnin parantaen suorituskykyä (LeCun ym. 2015). Tämä mahdollistaa erityisesti monimutkaisten ongelmanratkaisutehtävien suorittamisen, jotka ovat olleet aikaisemmilla koneoppimisen menetelmillä haastavia toteuttaa (LeCun ym., 2015). Lisäksi syväoppimista pystytään hyödyntämään tehokkaasti jäsentämättömän datan analysointiin, ja se tarjoaa koneoppimiseen verrattuna virheettömämpiä tuloksia (Jakhar & Kaur, 2020). Syväoppiminen onkin koneoppimisen alalaji (Jakhar & Kaur, 2020; Neittaanmäki ym., 2019). Syväoppiminen on tehostanut merkittävästi esimerkiksi puheentunnistusta, kuvantunnistusta, luonnollisen kielen prosessointia sekä konenäköä (Jordan & Mitchell, 2015; LeCun ym., 2015, 2021).

Syväoppimisessa yhdistyy algoritmit ja laskennalliset mallit, jotka muodostavat keinotekoisia neuroverkkoja (Jakhar & Kaur, 2020; Jordan & Mitchell, 2015; LeCun ym., 2015). Keinotekoiset neuroverkot imitoivat aivojen neuroverkkojen biologista rakennetta sekä toimintaa (Jakhar & Kaur, 2020). Ne muodostuvat kolmesta kerroksesta: syöttödataa vastaanottavasta syötekerroksesta,

dataprosessoinnista vastaavasta ulostuskerroksesta sekä näiden kerroksien välisistä piilokerroksista, joissa poimitaan tietoja datan luonteesta (Jakhar & Kaur, 2020; LeCun ym., 2015; Neittaanmäki ym., 2019). Nämä kerrokset taas rakentuvat neuroneista (Neittaanmäki ym., 2019).

Useita piiloverkkoja sisältäviä neuroverkkoja nimetään syviksi neuroverkoiksi, joihin syväoppiminen perustuu (Jakhar & Kaur, 2020; Neittaanmäki ym., 2019). Tällöin dataa käsitellään useilla piiloverkoilla, mikä mahdollistaa syväoppimiselle tyypillisen monimutkaisista tehtävistä suoriutumisen (Jakhar & Kaur, 2020). Syvät neuroverkot voivat sisältää miljoonia neuroneita, ja siten myös miljoonia parametrejä. Tästä syystä syväoppiminen vaatii suuren määrän opetusdataa (Neittaanmäki ym., 2019). Neuroverkkoja tulee lisäksi opettaa ja muokata, jotta ne saadaan toimimaan halutulla tavalla (Neittaanmäki ym., 2019). Koneoppimisen tavoin, syväoppimisessakin hyödynnetään ohjattua-, ohjaamatonta- ja vahvistusoppimista (LeCun ym., 2015). Kuviossa 1 Neittaanmäki ym. (2019, s. 30) ovat kuvanneet useita piilokerroksia sisältävän syvän neuroverkon rakennetta.



KUVIO 1 Syvän neuroverkon rakenne (Neittaanmäki ym., 2019, s. 30)

4 TEKOÄLYN ROOLI DISINFORMAATION LEVIÄMISESSÄ SOSIAALISESSA MEDIASSA

Tekoälyjärjestelmät voivat lisätä disinformaation riskiä mahdollistaen väärennetyn sisällön luomisen sekä tämän tehokkaan kohdennetun levittämisen internetissä (Bontridder & Poullet, 2021). Bontridder ja Poullet (2021) käsittelevät tutkimuksessaan verkossa esiintyviä tekoälyyn perustuvia disinformaation leviämisen ja torjumisen menetelmiä. He mainitsevat disinformaation levitykseen käytettävistä tekoälytekniikoista seuraavat: syväväärennökset, sosiaaliset botit ja mikrokohdentaminen. Näistä syväväärennökset mahdollistavat disinformaation luomisen, kun taas sosiaaliset botit ja mikrokohdentaminen tehostavat disinformaation leviämistä (Bontridder & Poullet, 2021). Tekoälypohjaista disinformaatiota käsitellään tässä tutkielmassa Bontridderin ja Poulletin (2021) rajauksen mukaisesti keskittyen syväväärennöksiin, sosiaalisiin botteihin sekä mikrokohdentamiseen. Tässä luvussa tutkitaan, kuinka tekoäly tukee näiden tekniikoiden toimivuutta ja edistää disinformaation leviämistä sosiaalisessa mediassa. Luvun ensimmäisessä alaluvussa käsitellään syväväärennöksiä, toisessa alaluvussa sosiaalisia botteja ja kolmannessa mikrokohdentamista.

4.1 Syväväärennökset

Syväväärennöksillä tarkoitetaan syväväärennösteknologioilla manipuloituja videoita, kuvia, audiota sekä tekstejä (Bontridder & Poullet, 2021). Syväväärennösten tuottaminen perustuu syväoppimiseen, erityisesti generatiiviseen kilpailevaan verkostoon, eli GAN-verkkoihin (engl. Generative Adversial Network) (Bontridder & Poullet, 2021; Preeti ym., 2023). Syväväärennöksiä saadaan aikaan kahden tekoälyalgoritmin toimiessa yhdessä GAN-verkossa, jolloin GAN-verkot luovat uutta dataa jo olemassa olevien tietojoukkojen perusteella (Bontridder & Poullet, 2021). Tällöin esimerkiksi kuvaa manipuloitaessa tekoälyllä analysoidaan tuhansia kuvia, joiden perusteella tekoäly pystyy luomaan uuden erilaisen kuvan samasta ilmiöstä (Bontridder & Poullet, 2021). Tätä teknologiaa voidaan hyödyntää erityyppisiin syväväärennöksiin, kuten kuviin,

videoihin, audioon ja tekstiin. Mirskyn ja Leen (2021) määritelmän mukaan syvävääreennökset ovatkin syvillä neuroverkoilla luotua uskottavaa mediaa. Syvävääreennösteknologiat mahdollistavat median manipuloinnin sekä kokonaan uuden sisällön luomisen (Bontridder & Poullet, 2021; Karnouskos, 2020).

Syvävääreennöksiä voidaan käyttää sisällön kuluttajaa vastaan kahdella tavalla: väärennettyä sisältöä voidaan pyrkiä esittämään aitona tai aidon sisällön voidaan väittää olevan väärennettyä (Kertysova, 2018). Syvävääreennöksillä voidaan pyrkiä vaikuttamaan yleiseen mielipiteeseen tai manipuloimaan todellista maailmaa (Karnouskos, 2020). Lisäksi syvävääreennökset voivat aiheuttaa luottamuksen puutetta informaatiota ja mediaa kohtaan (Karnouskos, 2020).

Shu ym. (2020) käsittelevät tutkimuksessaan sosiaalisessa mediassa esiintyvää disinformaatiota, painottaen syvävääreennettyjen kuvien ja videoiden olevan tyypillinen disinformaation esiintymismuoto tällaisilla alustoilla. Myös Karnouskos (2020) ja Mitra ym. (2021) korostavat tutkimuksissaan syvävääreennettyjen videoiden leviämisen olevan sosiaalisessa mediassa yleistä. Sillä laajan yleisön tavoittelu sosiaalisessa mediassa on helppoa, tekee se siitä otollisen paikan syvävääreennöksien levittämiselle (Karnouskos, 2020). Lisäksi sosiaalisen median alustat ovat alttiita väärinkäytöksille, kuten esimerkiksi kunnianloukkauksille ja kiristämiselle (Mitra ym., 2021).

Kertysovan (2018) mukaan onkin tyypillistä, että syvävääreennösteknologioilla pyritään vääristelemään siinä esiintyvän henkilön sanoja tai tekoja. Esimerkiksi syvävääreennetyille videoille tyypillistä on kasvojen, puheen tai tunteiden manipulointi (Mitra ym., 2021). Tällaisilla videoilla voi olla poliittisia, sosiaalisia ja emotionaalisia vaikutuksia niin yksilön tasolla kuin yhteiskunnallisesti (Mitra ym., 2021). Karnouskos (2020) painottaakin poliittisten videoiden manipuloinnin olevan hyvin tyypillinen syvävääreennösten esiintymismuoto.

Syvävääreennöksiin vaadittavat teknologiat ovat yksinkertaisia ja saatavilla laajasti (Karnouskos, 2020). Syvävääreennösten tuottaminen onkin nykyisin helppoa ja edullista jopa ei-valtiollisesti (Kertysova, 2018). Esimerkiksi sovellus FaceApp mahdollistaa väärennettyjen kuvien ja videoiden tuottamisen (Preeti ym., 2023).

Syvävääreennösteknologioita käsitellessä tulee huomioida, että niitä voidaan kaikesta huolimatta käyttää myös hyviin tarkoituksiin (Mitra ym., 2021). Esimerkiksi viihdearvoa luovat ei-haitalliset syvävääreennökset ovat myös sosiaalisessa mediassa yleisiä (Karnouskos, 2020)

4.2 Sosiaaliset botit

Hajli (2022), Bontridder ja Poullet (2021), Assenmacher ym. (2020) ja Shu ym. (2020) painottavat sosiaalisten bottien käyttöä sosiaalisessa mediassa disinformaation levittämisen keinona. Sosiaaliset botit ovatkin sosiaalisen median alustoilla toimivia käyttäjätilejä, joiden toiminta imitoi ihmisten kaltaista käyttäytymistä (Assenmacher ym., 2020; Bontridder & Poullet, 2021; Shu ym., 2020). Yksinkertaiset sosiaaliset botit automatisoivat esimerkiksi tykkäyksiä tai pos-

tausten jakamisia (Bontridder & Poullet, 2021). Tällaisia sosiaalisia botteja ei kuitenkaan Howardin ym. (2018) mukaan voida pitää älykkäinä. Kuitenkin erityisesti automatisoitujen, monimutkaisiin tehtäviin, kuten sisällöntuotantoon kykenevien bottien toiminta perustuu tekoälyteknologioihin, kuten koneoppimiseen (Assenmacher ym., 2020; Hajli, 2022). Älykkäät sosiaaliset botit pystyvät tarkkailemaan toimintaympäristöään ja vaikuttamaan ja reagoimaan siinä tietyn tavoitteen saavuttamiseksi (Howard ym., 2018). Tällaisia itsenäisesti toimivia botteja pidetään kaikista vaarallisimpia, niiden pystyessä tuottamaan tiettyyn kontekstiin liittyvää sisältöä automaattisesti (Assenmacher ym., 2020; Bontridder & Poullet, 2021).

Botit ovat yksinkertaistaen tietokonealgoritmeja (Aïmeur ym., 2023; Howard ym., 2018). Assenmacher ym. (2020) laajentavat tätä näkökulmaa painottaen sosiaalisten bottien toiminnan perustuvan infrastruktuuriin, jossa yhdistyy sosiaalisen median profiili ja automaatio, joka perustuu alustojen kaukoohjaukseen, sovellusliittymiin ja algoritmeihin, jotka ohjaavat botin toimintaa (Assenmacher ym., 2020). Ihminen voikin hallita botteja palvelimelta ja muodostaa yhteyden esimerkiksi Twitterin (nykyinen X) ohjelmointirajapintaan (API), joka mahdollistaa botin toiminnan alustalla (Howard ym., 2018). Ohjausjärjestelmä taas ohjaa botin käyttäytymistä tekemällä päätöksiä ohjelmointirajapinnalta saadun tiedon perusteella (Howard ym., 2018). Tekoälybottien opetusdatana voidaan hyödyntää suuria tekstiaineistoja, kuten romaaneja tai sosiaalisen median viestejä (Assenmacher ym., 2020). Lisäksi Assenmacher ym. (2022) mainitsevat joissakin tekoälypohjaisissa boteissa hyödynnettävän toistuvia neuroverkkoja (engl. recurrent neural networks) sekä LSTM-arkkitehtuureita (engl. long short term memory architecture). Orabi ym. (2020) painottavat haitallisten bottien takana olevan yleensä ihminen, ja botteja on mahdollista ostaa myös verkosta (Assenmacher ym., 2020).

Haitalliset sosiaaliset botit levittävät siis sosiaalisessa mediassa disinformaatiota vuorovaikuttamalla postausten kanssa lisäten niiden näkyvyyttä tai luomalla itsenäisesti tiettyä näkökulmaa puoltavaa sisältöä (Bontridder & Poullet, 2021). Botteja voidaankin käyttää sosiaalisen median alustoilla yleisen mielipiteen manipulointiin, poliittisten mielipiteiden muokkaamiseen sekä muun disinformaation levittämiseen (Hajli ym., 2022). Lisäksi botteja käytetään sosiaalisessa mediassa valeuutisten levittämiseen (Aïmeur ym., 2023; Hajli ym., 2022; Shu ym., 2020). Sosiaalisia botteja on hyödynnetty Twitterissä esimerkiksi vuonna 2016 Yhdysvaltain presidentinvaalien aikaan pyrkimyksenä vaikuttaa vaalituloksiin (Howard ym., 2018; Orabi ym., 2020; Shu ym., 2020). Botteja on esiintynyt Twitterissä huomattava määrä myös esimerkiksi Brexit-kansanäänestyksen aikana (Bastos & Mercea, 2018). Useat tutkimukset, kuten Hajli ym. (2022) ja Orabi ym. (2020) painottavatkin bottien esiintymistä Twitterissä, ja esimerkiksi vuonna 2017 arviolta 9–15 % Twitterin käyttäjistä käyttäytyi bottimaisesti (Shu ym., 2020).

Tutkijat ovat erimielisiä disinformaatiota levittävien tekoälybottien ilmenemisestä sosiaalisessa mediassa. Assenmacherin ym. (2020) mukaan tekoälybottien hyödyntäminen sosiaalisessa mediassa on vähäistä. Howard ym. (2018)

taas painottavat useiden Twitterissä esiintyvien bottien olevan älykkäitä. Myös Bontridder ja Poullet (2021) painottavat tekoälypohjaisten bottien roolia disinformaation levittäjinä sosiaalisessa mediassa.

Sosiaalisia botteja tarkastellessa on tärkeää huomioida, että niillä voi olla myös ei-haitallisia käyttötarkoituksia (Orabi ym., 2020). Lisäksi tulee ottaa huomioon, että botit saattavat levittää myös misinformaatiota disinformaation sijasta (Hajli ym., 2022). Lisäksi haitallisista sosiaalisista boteista saatetaan käyttää asiayhteydestä riippuen nimitystä poliittiset botit (engl. political bots) (Assenmacher ym., 2020; Howard ym., 2018) tai pahantahtoiset botit (engl. malicious bots) (Hajli ym., 2022).

4.3 Mikrokohdentaminen

Bontridder ja Poullet (2021) mainitsevat useiden sosiaalisen median alustojen, kuten Facebookin ja Twitterin hyödyntävän data-analyysiin ja algoritmeihin pohjautuvaa mikrokohdentamista personoidun sisällön jakamisen välineenä (Bontridder & Poullet, 2021). Myös Shankar (2024) painottaa useiden sosiaalisen median alustojen hyödyntävän tekoälyalgoritmeja sisällön personointiin sekä mainonnan kohdentamiseen. Tekoälypohjainen mikrokohdentaminen onkin liiketoimintamalli, joka on alun perin luotu mainonnan kohdentamiseksi potentiaalisille kuluttajille (Bontridder & Poullet, 2021). Tällainen sosiaalisen median alustojen hyödyntämä mikrokohdentaminen mahdollistaa kuitenkin myös disinformaatiokampanjoiden kohdentamisen haavoittuville käyttäjäryhmille (Bontridder & Poullet, 2021).

Mikrokohdentaminen perustuu henkilökohtaisen datan keräämiseen, jonka perusteella ihmiset luokitellaan ryhmiin ja ryhmille jaetaan personoitua sisältöä (Dobber ym., 2019). Kertysovan määritelmän (2018) mukaan käyttäjäprofilointi ja mikrokohdentaminen perustuu koneoppimiseen, jonka avulla yksilölle voidaan kohdistaa disinformaatiota erittäin henkilökohtaisella tasolla ottaen huomioon esimerkiksi yksilön demografiset perustiedot, kuten ikä ja sukupuoli sekä psykometriset tiedot, kuten persoonallisuuden piirteet. Tämä mahdollistaa disinformaation kohdistamisen tehokkaasti ottaen huomioon yksilön haavoittuvuudet (Kertysova, 2018). Esimerkiksi Facebookissa on aikaisemmin ollut pseudotieteistä kiinnostuneiden käyttäjäluokka, jota mainostajat ovat voineet hyödyntää mainonnan kohdentamisessa (Bontridder & Poullet, 2021). Myös Shankar (2024) painottaa, että sosiaalisen median tekoälyalgoritmeja voidaan manipuloida disinformaation levittämiseksi.

Simchonin ym. (2023) Redditiin perustuvan tutkimuksen mukaan persoonallisuuden piirteet vaikuttavatkin sisällön kuluttamiseen viitaten psykometrisen mikrokohdentamisen uhkaan. Esimerkiksi poliittisessa mikrokohdentamisessa koneoppiminen mahdollistaa äänestäjien ominaisuuksien analysoinnin, minkä perusteella voidaan pyrkiä ennustamaan, ketkä äänestäjistä ovat alttiita muuttamaan äänestyspäätöksiään (Bennett & Gordon, 2021). Eräs esimerkki poliittisesta mikrokohdentamisesta on Yhdysvaltain presidentin vaalit vuonna

2016, jossa hyödynnettiin Cambridge Analytican psykometristä profilointia pyrkimyksenä vaikuttaa äänestäjiin ja siten vaalituloksiin (Kertysova, 2018). Walker ym. (2019) käsittelevät tutkimuksessaan vastaavaa ilmiötä Facebookissa. Simchon ym. (2023) painottavatkin huolta poliittista mikrokohdentamista kohtaan sosiaalisen median roolin kasvaessa poliittisessa ympäristössä (Simchon ym., 2023).

Kertysova lisää tekoälyn mahdollistavan myös sisällöntuotannon automatisoinnin sen yhdistyessä vahvasti mikrokohdentamiseen (Kertysova, 2018). Tällä tarkoitetaan automatisoitua sisällöntuotantoa, jossa sisältö räätälöidään ja kohdennetaan yksilöllisesti käyttäjän henkilökohtaisia piirteitä hyödyntäen (Kertysova, 2018). Karnouskos (2020) liittää tämän sosiaalisen median kontekstiin viitaten syväväärennöksiin. Karnouskosin mukaan tekoäly mahdollistaa automaattisen kuvien, videoiden ja tekstin tuottamisen digitaalisella alustalla esiintyvän datan perusteella esimerkiksi Facebookissa, Instagramissa tai Twitterissä. Tämän perustuu luonnollisen kielen prosessoinnin ja syväoppimisen kehittymiseen (Karnouskos, 2020). Luonnollisen kielen prosessoinnilla tarkoitetaan laskennallista ihmiskielen ymmärtämistä ja tuottamista, pitäen sisällään esimerkiksi konekääntämisen ja puheentunnistamisen (Hirschberg & Manning, 2015).

Lisäksi mikrokohdentaminen edistää sosiaalisessa mediassa myös jo aiemmin mainittujen suodatinkuplien ja kaikukammioiden muodostumista lisäten disinformaation riskiä (Bennett & Gordon, 2021). Kukin sosiaalisen median käyttäjä näkee siis erilaisen version todellisuudesta, sillä suosittelualgoritmit vaikuttavat siihen, millaista sisältöä käyttäjälle syötetään (Bontridder & Poullet, 2021).

Myös mikrokohdentamista käsiteltäessä tulee ottaa huomioon, että se voi edistää disinformaation lisäksi myös misinformaation leviämistä (Bontridder & Poullet, 2021; Shankar, 2024). Taulukossa kaksi esitetään yhteenveto syväväärennösten, sosiaalisten bottien ja mikrokohdentamisen roolista disinformaation levittämisessä sosiaalisessa mediassa.

TAULUKKO 2 Teknologioiden rooli disinformaation levittämisessä

Teknologia	Tarkoitus	Tavoite
Syväväärennökset	Uskottavan disinformaatiota sisältävän median tuottaminen	Harhaanjohtaminen ja manipulointi
Sosiaaliset botit	Disinformaation kanssa vuorovaikuttaminen ja sisällöntuotanto	Disinformaation leviämisen tehostaminen sen näkyvyyttä parantamalla
Mikrokohdentaminen	Disinformaation kohdentaminen yksilöllisesti	Disinformaation leviämisen tehostaminen

5 TEKOÄLYN ROOLI DISINFORMAATION TORJUMISESSA SOSIAALISESSA MEDIASSA

Tässä luvussa käsitellään tekoälyn roolia disinformaation torjumisessa sosiaalisessa mediassa. Tämän luvun ensimmäisessä alaluvussa käsitellään tekoälyn roolia disinformaation torjumisessa sosiaalisessa mediassa. Luvun toisessa alaluvussa taas käsitellään lyhyesti tekoälyn käytön haasteita disinformaation torjunnassa.

5.1 Tekoälyn rooli disinformaation torjumisessa

Vaikka tekoäly on edistänyt disinformaation levittämisen keinoja, voidaan sitä hyödyntää myös disinformaation tunnistamiseen ja torjumiseen (Kertysova, 2018). Sosiaalisessa mediassa manuaalinen faktantarkistus on usein tehotonta datan määrän ollessa suurta (Rastogi & Bansal, 2022). Tämä luo tarvetta automaattisesti toimiville tekoälypohjaisille faktantarkistusjärjestelmille (Rastogi & Bansal, 2022). Kertysovan (2018) mukaan tekoälyyn pohjautuvien faktantarkistusjärjestelmien kehittämisprojektit sosiaalisessa mediassa ovatkin lisääntyneet 2010-luvulla. Tällaisilla järjestelmillä tarkoitetaan menetelmiä, jotka automaattisesti tunnistavat, tarkastavat sekä korjaavat median sisältöjä (Kertysova, 2018). Suurin osa tällaisista järjestelmistä pohjautuu Kertysovan mukaan koneoppimiseen (2018). Useat sosiaalisen median alustat, kuten Facebook ja Twitter, hyödyntävätkin koneoppimisalgoritmeja trollien, bottien ja muun haitallisen tiedon, kuten terrorismin tunnistamiseen sekä poistamiseen alustoilla (Kertysova, 2018). Tekoälypohjainen faktantarkistus voikin parhaassa tapauksessa parantaa sosiaalisen median alustojen turvallisuutta (Shankar, 2024) sekä vähentää ihmismoderaattorien stressiä (Kertysova, 2018). Tämän luvun alaluvuissa keskitytään käsittelemään tarkemmin tekstimuotoisen disinformaation, syvävääreennettyjen kuvien ja videoiden, sosiaalisten bottien sekä mikrokohdentamisen torjuntamenetelmiä.

5.1.1 Tekstimuotoisen disinformaation torjuminen

Eräs esimerkki tekstimuotoisen disinformaation tunnistamisesta sosiaalisessa mediassa on Rastogin ja Bansalin (2022) esittämä koneoppimismetodi, joka luokittelee Twitterin twiitteja informaatiotyypeittäin disinformaatioksi, satiiriksi sekä oikeelliseksi tiedoksi tiedon aitouden ja levittämisen tarkoituksen perusteella (Rastogi & Bansal, 2022). Menetelmä perustuu siihen, että väärennetty teksti on eri tavoin kirjoitettu oikeelliseen tekstiin verrattuna. Menetelmässä väärennetyn tekstin tunnistamiseksi analysoitiin muun muassa sana- ja merkkimäärää, uudelleentwiittausten määrää sekä sanojen tunnetiloja (Rastogi & Bansal, 2022). Menetelmä koostui twiittien poimimisesta, twiittien ominaisuuksien erottelusta ohjattuja koneoppimisluokittelijoita käyttäen sekä merkittävien piirteiden erottelusta ja luokittelusta ANOVA-tilastotestejä hyödyntäen. Rastogin ja Bansalin (2022) menetelmä toimi tehokkaasti informaation luokittelussa ensemble-koneoppimismenetelmää hyödyntäen, jossa erilaisia luokittelijoita yhdistellään tarkoituksena parantaa tulosten tarkkuuta (Rastogi & Bansal, 2022). Tällaista metodia voitaisiin mahdollisesti käyttää Twitterissä (nykyinen X) harhaanjohtavien twiittien merkitsemiseksi.

Kun Rastogi ja Bansal (2022) painottavat ohjattua koneoppimismenetelmää disinformaation tunnistamisessa sosiaalisessa mediassa, Yang ym. (2019) taas painottavat ohjaamattomaan oppimiseen perustuvaa valeuutisten tunnistusmenetelmää sosiaalisessa mediassa. Yangin ym. (2019) menetelmässä hyödynnettiin sosiaalisen median käyttäjien sitoutumisdataa valeuutisten tunnistamisen keinona. Menetelmä on siis hieman erilainen Rastogin ja Bansalin (2022) menetelmään verrattuna, joka perustui tekstin analysointiin.

Lisäksi Maathuis ja Godschalk (2023) ehdottavat tutkimuksessaan syväoppimismenetelmää tekstimuotoisen disinformaation tunnistamiseksi sosiaalisessa mediassa. Heidän metodinsa perustuu kahteen syväoppimistekniikkaan: konvulaationeuroverkkoihin (CNN) sekä kaksisuuntaisiin enkooderiesityksen muuntajiin (BERT), jotka hyödyntävät tunneanalyysiä disinformaation tunnistamiseksi koronapandemian aikaisista twiiteista (Maathuis & Godschalk, 2023).

Erilaiset kone- ja syväoppimismenetelmät voivat siis mahdollistaa tekstimuotoisen disinformaation torjumisen sosiaalisessa mediassa. Lisäksi sosiaalisen median alustat voivat pyrkiä ehkäisemään disinformaatiota muuttamalla algoritmejaan siten, että ne eivät suosi disinformaatiota, vaan pyrkivät kumoamaan sitä (Kertysova, 2018).

5.1.2 Syväväärennösten torjuminen

Mirsky ja Lee (2021) käsittelevät tutkimuksessaan syväväärennösten tunnistamista yleisluonteisesti. He jakavat syväväärennösten tunnistamisen artefaktien tunnistamiseen ja suuntaamattomiin lähestymistapoihin. Artefaktien tunnistaminen voi perustua esimerkiksi ympäristön epäjohdonmukaisuuksien (esim. valaistus) tai oikeudellisten artefaktien (esim. GAN-verkon sormenjäljet) havaitsemiseen sekä käyttäytymisen poikkeamien tai fysiologisten signaalien puutteen analysointiin (esim. silmien räpyttelyn epäjohdonmukaisuus) (Mirksy

& Lee, 2021). Suuntaamattomia lähestymistapoja taas ovat luokittelu sekä poikkeamien havaitseminen (Mirksy & Lee, 2021).

Kun kuva tai video ladataan sosiaaliseen median alustalle, se tyypillisesti pakataan (Mitra ym., 2021). Shu ym. (2020) painottavatkin, että sosiaalisessa mediassa jaettujen syvävääärennettyjen kuvien tunnistamismenetelmien tulisi olla yhteensopivia pakattujen kuvien kanssa. Shu ym. (2020) mainitsevat esimerkiksi Xception-verkon toimivan tehokkaasti pakattujen StyleGAN-verkolla luotujen syvävääärennettyjen kuvien tunnistamisessa. Preeti ym. (2023) taas käsittelevät julkaisussaan GAN-verkkoihin pohjautuvaa syvävääärennettyjen kuvien tunnistusmenetelmää sosiaalisessa mediassa. Tutkimuksen perusteella syvät konvoluutio-GAN-verkot mahdollistavat väärennettyjen kuvien tunnistamisen jopa suhteellisen rajoitettujen datajoukkojen perusteella. (Preeti ym., 2023.)

Mitra ym. (2021) taas käsittelevät syvävääärennettyjen videoiden tunnistamista sosiaalisessa mediassa. Heidän mukaansa syvävääärennettyjen videoiden tunnistamisessa tyypillistä on esimerkiksi päänasennon, silmien räpyttelyn, kasvojen alueen ja hampaiden artefaktien analysointi (Mitra ym., 2021). Lisäksi voidaan pyrkiä analysoimaan videossa tai audiossa esiintyviä tunteita (Mitra ym., 2021). Mitran ym. (2021) menetelmä perustuu konvoluutiokeinoverkkoon Xception sekä luokitinverkkoon. Menetelmässä syvävääärennetyt videot tunnistetaan visuaalisten artefaktien perusteella (Mitra ym., 2021). Tällöin konvoluutiokeinoverkko Xception toimii ominaisuuspoimijana, jonka jälkeen ominaisuudet luokitellaan luokitinverkossa. Lisäksi metodissa hyödynnettiin avainvideokehysten poimintatekniikkaa, jolla pyrittiin vähentämään analysoitavien kehysten määrää yhtä videota kohden. (Mitra ym., 2021.) Mitran ym. (2021) menetelmä tarjoaa laskennallisesti tehokkaan tavan tunnistaa syvävääärennettyjä pakattuja videoita sosiaalisessa mediassa ilman laajaa opetusdataa. Tämä mahdollistaa menetelmän käytön myös laitteissa, joissa muistin käyttö on rajattua (2021). Tällöin videon alkuperän ja aitouden selvittäminen voisi olla mahdollista esimerkiksi älypuhelimella (Mitra ym., 2021).

Näiden tutkimusten perusteella syvävääärennettyjen kuvien ja videoiden torjunta sosiaalisessa mediassa perustuu tyypillisesti syväoppimismenetelmiin, jotka hyödyntävät syvävääärennöksille tyypillisten artefaktien analysointia. Lisäksi tällaisten tunnistusmenetelmien on tärkeää olla yhteensopivia pakatussa muodossa olevien kuvien ja videoiden kanssa.

5.1.3 Sosiaalisten bottien torjuminen

Kertysovan (2018) mukaan koneoppimisratkaisut ovat osoittautuneet tehokkaiksi bot-tilien tunnistamisessa ja merkitsemisessä. Myös Assenmcher ym. (2020) ja Shu ym. (2020) painottavat bottien tunnistamisessa koneoppimisen menetelmiä. Shun ym. (2020) mukaan eräs tapa tunnistaa sosiaalisia botteja on pyrkiä tunnistamaan eroja bottien ja ihmiskäyttäjien välillä. Twitterissä boteille tyypillisiä piirteitä ovat usein toistuvat URL-osoitteet, uudelleentwiittaukset, räjähdysmäinen twiittaaminen lyhyessä ajassa sekä useiden käyttäjien seuraaminen (Shu ym., 2020). Shun ym. (2020) mukaan näiden ominaisuuksien tunnis-

tamisessa voidaan hyödyntää useita koneoppimiseen pohjautuvia luokittelumenetelmiä, kuten neuroverkkoja, satunnaismetsiä sekä tukivektorikoneita (Shu ym., 2020).

Myös Orabi ym. (2020) tuovat artikkelissaan esille useita eri sosiaalisten bottien tunnistamiseen hyödynnettäviä koneoppimistekniikoita keskittyen Twitteriin. Heidän mukaansa ohjattu oppiminen on yleisimmin bottien tunnistamisessa hyödynnetty koneoppimisparadigma. Heidän esittämässä metodissa luokittelija opetetaan opetusdatan perusteella tunnistamaan sosiaaliset botit niille tyypillisten piirteiden perusteella (Orabi ym., 2020). Ohjatussa oppimisessa haasteeksi voi kuitenkin muodostua bottien nopea kehittyminen ja ajantasaisen opetusdatan puute (Orabi ym., 2020). He tuovatkin esille myös ohjaamattomaan oppimiseen pohjautuvan bottien tunnistusmetodin, jossa botit pyritään tunnistamaan klusteroimalla samankaltaisia käyttäjätilejä käyttäytymisen ja sisällöllisen analysoinnin perusteella (Orabi ym., 2020). Tällöin klusteroinnin perusteena voi olla esimerkiksi samankaltaisen sisällön jakaminen, hyvin aktiivinen postaaminen tai trendaavien hashtagien käyttäminen (Orabi ym., 2020). Myös puoliohjattua oppimista voidaan käyttää bottien tunnistamiseen, mutta tutkimusta siitä on vain vähän (Orabi ym., 2020).

Toinen esimerkki sosiaalisten bottien tunnistamisesta on Hajlin (2022) ehdottama menetelmä, jossa hyödynnetään toimijaverkkoteoriaa (engl. Actor-network theory). Hajlin tutkimuksessa hyödynnettiin 30 000 twiittiä, joiden piirteiden analysoinnissa ja luokittelussa hyödynnettiin tekstin louhintaa ja kolmea eri koneoppimisalgoritmia. Tämän jälkeen kuutta koneoppimismetodia sekä kahta syväoppimismetodia testattiin pyrkimyksenä erotella ja luokitella bottien ja ihmiskäyttäjien luomat twiitit toisistaan (Hajli ym., 2022). Parhaiten bottien ja ihmiskäyttäjien erottamisesta suoriutui syväoppimisen menetelmä Bi-LSTM (Hajli ym., 2022).

Kone- ja syväoppiminen tarjoavat siis erilaisia metodeja bottien tunnistamiseksi sosiaalisessa mediassa. Näiden tutkimuksien perusteella bottien tunnistaminen perustuu usein boteille ominaisten ominaisuuksien analysointiin kone- ja syväoppimismenetelmillä. Tekoälyratkaisujen lisäksi esimerkiksi kaksivaiheisen tunnistamisen käyttäminen voi rajoittaa ei-haluttujen bottien pääsyä sosiaalisen median alustoille (Hajli, 2022).

5.1.4 Mikrokohdentamisen torjuminen

Tutkimus haitallisen mikrokohdentamisen torjumisesta on toistaiseksi vähäistä. Sosiaalisen median alustat saavat taloudellista hyötyä mikrokohdentamisesta tarjoamalla käyttäjille personoitua sisältöä ja kohdennettua mainontaa (Shankar, 2024). Tämä saattaa vaikuttaa siihen, miksi mikrokohdentamisen uhkien torjumisesta on vain vähän tietoa.

Simchonin ym. (2023) mukaan haitallisen mikrokohdentamisen ehkäisykeinona saattaisi kuitenkin toimia mikrokohdentamisen algoritmien käänteinen rakentaminen siten, että käyttäjää voitaisiin varoittaa hänen kuluttaessaan mikrokohdennettua sisältöä. Tämä saattaisi auttaa sosiaalisen median käyttäjiä kuluttamaan sisältöä kriittisemmin.

Lisäksi on olemassa joitakin tekoälystä riippumattomia tapoja ehkäistä mikrokohdentamisen uhkia. Eräs tapa lisätä tietoisuutta mikrokohdentamisesta on parantaa käyttäjien ymmärrystä siitä (Lorenz-Spreen ym., 2021). Mikrokohdennetun mainonnan tunnistamista voidaankin mahdollisesti tehostaa esimerkiksi omaa persoonallisuutta tutkiskelemalla kyselylomakkeen avulla (Lorenz-Spreen ym., 2021). Toinen esimerkki tietoisuuden lisäämisestä on Facebookin ”Miksi näen tämän mainoksen?”-painike, jolla alusta pyrkii antamaan tietoa henkilökohtaisen datan käytöstä (Lorenz-Spreen ym., 2021). Tällaiset painikkeet tarjoavat kuitenkin vain pinnallista tietoa, ja tietoisuus henkilökohtaisen datan haitallisesta käytöstä saattaa johtaa mainoksien tehottomuuteen (Lorenz-Spreen ym., 2021).

Digitaalisilla alustoilla onkin velvollisuus kertoa henkilökohtaisen datan käytöstä käyttäjille (Lorenz-Spreen ym., 2021). Esimerkiksi Euroopan Unionin yleisessä tietosuojalaissa GDPR:ssä otetaan kantaa henkilökohtaisen datan keräämiseen ja käsittelyyn (Dobber ym., 2019; Kertysova, 2018). Algoritmien monimutkaisuus tuottaa kuitenkin ongelmia säädösten noudattamisessa (Kertysova, 2018). Taulukossa kolme vedetään yhteen kaikki luvussa 5.1 esitellyt disinformaation torjuntamenetelmät.

TAULUKKO 3 Disinformaation torjuntamenetelmien yhteenveto

Tutkimus	Tavoite	Hyödynnetty teknologia
Kertysova, 2018	Disinformaation torjuminen	Koneoppiminen
Rastogi & Bansal, 2022	Disinformaation torjuminen	Koneoppiminen
Yang ym., 2019	Disinformaation torjuminen	Koneoppiminen
Maathuis ja Godschalk, 2023	Disinformaation torjuminen	Syväoppiminen
Shu ym., 2020	Syvävääreennettyjen kuvien torjuminen	Syväoppiminen
Preeti ym., 2023	Syvävääreennettyjen kuvien torjuminen	Syväoppiminen
Mitra ym., 2021	Syvävääreennettyjen videoiden torjuminen	Syväoppiminen
Shu ym., 2020	Sosiaalisten bottien torjuminen	Koneoppiminen
Orabi ym., 2020	Sosiaalisten bottien torjuminen	Koneoppiminen
Hajli ym., 2022	Sosiaalisten bottien torjuminen	Koneoppiminen / syväoppiminen
Hajli ym., 2022	Sosiaalisten bottien torjuminen	Kaksivaiheinen tunnistautuminen
Simchon ym., 2023	Mikrokohdentamisen haittojen torjuminen	Algoritmien hyödyntäminen
Lorenz-Spreen ym., 2021	Mikrokohdentamisen haittojen torjuminen	Tietoisuuden lisääminen

5.2 Tekoälypohjaisen faktantarkistuksen haasteet

Vaikka tekoäly tarjoaakin mahdollisuuksia disinformaation tunnistamiseksi sosiaalisessa mediassa, täysin automaattinen faktantarkistus on vielä saavuttamatta (Kertysova, 2018). Kertysovan mukaan esimerkiksi vuonna 2018 Facebook on työllistänyt vielä 7500 ihmismoderaattoria alustan sisällön tarkistamiseen (2018). Myös Bontridder ja Poullet (2021) painottavat, että vaikka tekoälyllä voidaan tehokkaasti tunnistaa disinformaatiota, ei se kuitenkaan ole kykeneväinen arvioimaan tiedon laatua.

Tekoälyjärjestelmille tyypillistä onkin alttius väärille positiivisille ja negatiivisille tuloksille, mikä voi johtaa informatiivisen tiedon luokitteluun disinformaatioksi (Kertysova, 2018). Lisäksi Kertysovan mukaan tekoälypohjaiset faktantarkistusjärjestelmät ovat usein tehottomia tunnistamaan esimerkiksi sar-

kasmia tai ironiaa aiheuttaen väärinymmärryksiä informaation luonteesta. Lisäksi hän painottaa, että koneoppimiseen perustuvat tekoälyjärjestelmät voivat edistää ennakkoluulojen ja syrjinnän leviämistä, jos järjestelmä on koulutettu puutteellisella datalla. Koneoppimiseen sekä keinotekoisiin neuroverkkoihin pohjautuvien tekoälyjärjestelmien monimutkaisuus voi lisäksi vaikeuttaa tekoälyn antamien lopputulemien selittämistä. (Kertysova, 2018.) Tällöin voi olla kyseenalaista, toimiiko esimerkiksi informaation laadun tunnistaminen tarkoituksenmukaisella tavalla (Kertysova, 2018). Lisäksi Kertysovan mukaan tekoälypohjaisten faktantarkistusjärjestelmien automatisoituminen saattaa tulevaisuudessa johtaa siihen, että tekoäly alkaa tehdä yhä itsenäisempiä päätöksiä. Tämä saattaa vähentää ihmisen valvontaa ja johtaa siihen, että tekoäly manipuloi tietoja itsenäisesti (Kertysova, 2018).

Kertysova painottaakin, että ihmisen tulisi tarkistaa koneen tekemät päätökset tiedon oikeellisuuden ja oikeudenmukaisuuden varmistamiseksi. Hän painottaa, että digitaaliset ratkaisut eivät yksinään pysty torjumaan disinformaatiota vaan ongelma on myös inhimillinen. Esimerkiksi media- ja digilukutaidon harjoittaminen onkin teknologisten ratkaisujen ohella hyvin tärkeää disinformaation leviämisen ehkäisyssä. (Kertysova, 2018.)

6 YHTEENVETO

Tämän tutkielman tavoitteena oli ymmärtää disinformaation leviämiseen ja torjumiseen vaikuttavien tekoälyteknologioiden toimintamekanismeja, ja tarkastella niiden kaksinaista roolia disinformaation levittäjinä ja torjumisen mahdollistajina sosiaalisen median kontekstissa. Tutkimuksessa vastattiin tutkimuskysymykseen ”Miten tekoäly vaikuttaa disinformaation leviämiseen sosiaalisessa mediassa, ja millaisia mahdollisuuksia se voi tarjota tämän ehkäisemiseksi?”. Aiempaa tutkimusta aiheesta löytyi rajallisesti, mikä loi tarpeen ilmiön tutkimiselle.

Tutkimus toteutettiin kuvailevana kirjallisuuskatsauksena, jossa hyödynnettiin 42 lähdettä. Tutkielman aineisto kerättiin Jykdok-, Google Scholar-, IEEE Xplore-, Scopus- sekä Keenious-tietokannoista. Aineistoon pyrittiin valitsemaan Julkaisufoorumissa julkaistuja vertaisarvioituja lähteitä.

Tutkielmassa havaittiin, että tekoälyteknologioiden kehittyminen on vaikuttanut disinformaation leviämiseen sosiaalisessa mediassa. Tekoälyyn pohjautuvan disinformaation käsittely rajattiin tutkimuksessa Bontridderin ja Poulletin (2021) määritelmän mukaisesti syvävääreennöksiin, sosiaalisiin botteihin ja mikrokohdentamiseen. Tutkimustulokset osoittivat erityisesti syvävääreennettyjen kuvien ja videoiden (Karnouskos, 2020; Mitra ym., 2021; Shu ym., 2020), sosiaalisten bottien (Bontridder & Poullet, 2021; Hajli, 2022; Shu ym., 2020) sekä mikrokohdentamisen (Bontridder & Poullet, 2021) olevan disinformaation levittämiseen hyödynnettyjä tekoälyyn perustuvia teknologioita sosiaalisessa mediassa. Tutkielmassa todettiin syväoppimiseen perustuvien (Bontridder & Poullet, 2021; Mirsky & Lee, 2021) syvävääreennösteknologioiden mahdollistavan disinformaatiota sisältävän median tuottamisen, ja tämän levittämisen sosiaalisessa mediassa. Lisäksi havaittiin, että sosiaaliset botit voivat lisätä disinformaation leviämistä toimimalla alustoilla aktiivisesti ja tarkoituksellisesti tiettyä päämäärää tukien. Sosiaalisten bottien teknologiassa painottuivat erilaiset koneoppimismenetelmät (Assenmacher ym., 2020; Hajli, 2022; Howard ym., 2018). Myös mikrokohdentamisessa korostui koneoppimisen hyödyntäminen (Bennett & Gordon, 2021; Bontridder & Poullet, 2021; Shankar, 2024). Tutkimustulokset osoittivat mikrokohdentamisen mahdollistavan disinformaatiokampanjoiden

tehokkaan kohdentamisen yksilön demografisia ja psykometrisiä tietoja hyväksikäyttäen. Tätä tutkielmaa tarkasteltaessa on tärkeää huomioida, että näillä teknologioilla on myös ei-haitallisia käyttötarkoituksia.

Toisaalta tutkielmassa havaittiin tekoälyn mahdollistavan myös disinformaation torjumisen sosiaalisessa mediassa. Tutkielmassa otettiin kantaa tekstimuotoisen disinformaation, sosiaalisten bottien, syvävääreennösten sekä mikrokohdentamisen tekoälypohjaisiin torjuntamenetelmiin sosiaalisen median kontekstissa. Tekstimuotoisen disinformaation torjumisessa sosiaalisessa mediassa korostuivat erilaiset kone- ja syväoppimismetodit, jotka mahdollistivat harhaanjohtavan ja oikeellisen tiedon erottamisen toisistaan niille tyypillisten ominaisuuksien perusteella. Myös sosiaalisten bottien torjunta perustui kone- ja syväoppimismenetelmiin, jotka pyrkivät tunnistamaan bottitilejä analysoimalla eroja ihmis- ja bottikäyttäjien välillä. Syvävääreennösten torjunnassa taas painottuivat erilaiset syväoppimismenetelmät. Syvävääreennösten tunnistaminen perustui tyypillisesti väärennöksille ominaisten artefaktien tunnistamiseen. Lisäksi tunnistusmetodin oli tärkeää olla yhteensopiva sosiaalisessa mediassa esiintyvien pakatussa muodossa olevien kuvien ja videoiden kanssa. Mikrokohdentamisen uhkien tekoälypohjaisesta torjumisesta löytyi hyvin vähän tutkimusta, ja sen ehkäisyssä korostuivat inhimilliset torjuntatavat. Algoritmeja voidaan kuitenkin mahdollisesti hyödyntää käyttäjien tiedottamiseksi heidän altistuessaan mikrokohdennetulle sisällölle.

Vaikka tekoälyteknologiat voivat tehostaa disinformaation tunnistamista, liittyy niiden käyttämiseen faktantarkistuksessa myös edelleen huomattavia haasteita (Kertysova, 2018). Vaikka tekoäly on tehokas tunnistamaan disinformaatiota, se ei kuitenkaan ole kykenevä arvioimaan tiedon laatua (Bontridder & Poullet, 2021). Teknologisten ratkaisujen ohella myös inhimilliset torjuntatavat korostuvatkin edelleen disinformaation torjunnassa (Kertysova, 2018).

Tämä tutkielma korostaa tekoälyn olevan sekä uhka että mahdollisuus disinformaation leviämiselle sosiaalisessa mediassa. Tekoälyä ei siis voida pitää yksiselitteisesti hyvänä tai pahana, vaan sen vaikutukset disinformaation leviämisessä riippuvat pitkälti sen käyttötavasta. Vaikka tekoäly voi edistää disinformaation leviämistä, se voi parhaimmillaan myös parantaa sosiaalisen median alustojen turvallisuutta. Tekoälyn käytön sääntelyssä tulisikin ottaa huomioon tekoälyn kaksinainen rooli disinformaation torjujana ja levittäjänä.

Tutkimustulokset voivat lisätä digi- ja medialukutaitoa lisäten ymmärrystä informaation vaaroista sosiaalisessa mediassa. Lisäksi tutkimus voi luoda raameja tekoälyteknologioiden vastuulliselle käytölle, väärinkäytöksien estämiselle ja mahdolliselle yksilöä suojaavan lainsäädännön kehittämiselle.

Tässä tutkielmassa ei tarkasteltu tekoälypohjaisen faktantarkistuksen eettisiä ongelmia, kuten sananvapauden rajoittumista tai yksityisyydensuojaa. Laajempi tutkimusaineisto voisi puolestaan tuottaa luotettavampia tutkimustuloksia. Lisäksi disinformaation käsitteen vakiintumattomuus asetti haasteita tutkimusprosessin toteuttamiselle. Tutkimuksen esimerkkitapauksista huomattava osa keskittyi Twitteriin. On tärkeää huomata, että Twitter tunnetaan nykyisin nimellä X, ja yrityksen omistajuudessa on tapahtunut muutoksia. Olisi-

kin tärkeää tutkia, miten nämä muutokset ovat vaikuttaneet disinformaation leviämiseen ja torjumiseen palvelussa. Tekoälyn vaikutuksia disinformaation leviämisessä tulisi selvittää tarkemmin myös muilla sosiaalisen median alustoilla. Jatkotutkimusaiheeksi ehdotan sosiaalisen median palvelu X:n merkityksen tutkimista disinformaation levitysalustana.

LÄHTEET

Abbass, H. (2021). Editorial: What is Artificial Intelligence? *IEEE Transactions on Artificial Intelligence*, *2*(2), 94–95. IEEE Transactions on Artificial Intelligence. https://doi.org/10.1109/TAI.2021.3096243

Ahmad, N., Milic, N., & Ibahrine, M. (2021). Data and Disinformation. *Computer (Long Beach, Calif.)*, *54*(7), 105–110. https://doi.org/10.1109/MC.2021.3074261

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, *13*(1), 1–30. https://doi.org/10.1007/s13278-023-01028-5

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media + Society*, *6*(3), 1–14. https://doi.org/10.1177/2056305120939264

Bastos, M. & Mercea, D. (2018). The public accountability of social platforms: Lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). https://doi.org/10.1098/rsta.2018.0003

Becker, H. (1949). The Nature and Consequences of Black Propaganda. *American Sociological Review*, *14*(2), 221–235. https://doi.org/10.2307/2086855

Bennett, C. J., & Gordon, J. (2021). Understanding the "Micro" in Political Micro-Targeting: An Analysis of Facebook Digital Advertising in the 2019 Federal Canadian Election. *Canadian Journal of Communication*, *46*(3), 431–459. https://doi.org/10.22230/cjc.2021v46n3a3815

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy*, *3*, e32. https://doi.org/10.1017/dap.2021.20

Bradshaw, S., & DeNardis, L. (2024). Technical infrastructure as a hidden terrain of disinformation. *Journal of Cyber Policy*, 1–17. https://doi.org/10.1080/23738871.2024.2419010

Carr, C. T., & Hayes, R. A. (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

Dobber, T., Ó Fathaigh, R., & Zuiderveen Borgesius, F. J. (2019). The regulation of online political micro-targeting in Europe. *Internet Policy Review*, *8*(4), 1–20. https://doi.org/10.14763/2019.4.1440

Fallis, D. (2015). What Is Disinformation? *Library Trends*, *63*(3), 401–426.

Fetzer, J. H. (2004). Disinformation: The Use of False Information. *Minds and Machines*, *14*(2), 231–240. https://doi.org/10.1023/B:MIND.0000021683.28604.5b

Guth, D. W. (2009). Black, White, and Shades of Gray: The Sixty-Year Debate Over Propaganda versus Public Diplomacy. *Journal of Promotion Management*, *14*(3–4), 309–325. https://doi.org/10.1080/10496490802624083

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social Bots and the Spread of Disinformation in Social Media: The Challenges of Artificial Intelligence. *British Journal of Management*, *33*(3), 1238–1253. https://doi.org/10.1111/1467-8551.12554

Hirschberg, J. & Manning, C. D. (2015). Advances in natural language processing. *Science (American Association for the Advancement of Science)*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Howard, P. N., Woolley, S., & Calo, R. (2018). Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration. *Journal of Information Technology & Politics*, *15*(2), 81–93. https://doi.org/10.1080/19331681.2018.1448735

Jakhar, D., & Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: Definitions and differences. *Clinical and Experimental Dermatology*, *45*(1), 131–132. https://doi.org/10.1111/ced.14029

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Karnouskos, S. (2020). Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138–147. IEEE Transactions on Technology and Society. https://doi.org/10.1109/TTS.2020.3001312

Kertysova, K. (2018). Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered. *Security and Human Rights*, *29*(1–4), 55–81. https://doi.org/10.1163/18750230-02901005

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature (London)*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Bengio, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58–65. https://doi.org/10.1145/3448250

Lorenz-Spreen, P., Geers, M., Pachur, T., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Boosting people's ability to detect microtargeted

advertising. *Scientific Reports*, *11*(1), 15541. https://doi.org/10.1038/s41598-021-94796-z

Maathuis, C., & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection. *International Conference on Cyber Warfare and Security*, *18*(1), 237–245. https://doi.org/10.34190/iccws.18.1.951

McCarthy, J. (2004). *What is artificial intelligence*. http://cse.unl.edu/~choueiry/S09-476-876/Documents/whatisai.pdf

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, *54*(1), 7:1-7:41. https://doi.org/10.1145/3425780

Mitra, A., Mohanty, S. P., Corcoran, P., & Kougianos, E. (2021). A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame Extraction. *SN Computer Science*, *2*(2), 98. https://doi.org/10.1007/s42979-021-00495-x

Neittaanmäki, P., Tuominen, H., Niinimäki, E., Pölönen, I., Rautiainen, I., Äyrämö, S., Ruohonen, T., Nyrhinen, R., Ojalainen, A., Vähäkainu, P., & Äyrämö, S.-M. (2019). *Tekoälyn perusteita ja sovelluksia*. https://jyx.jyu.fi/handle/123456789/64975

Obar, J. A., & Wildman, S. (2015). Social media definition and the governance challenge: An introduction to the special issue. *Telecommunications Policy*, *39*(9), 745–750. https://doi.org/10.1016/j.telpol.2015.07.014

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management*, *57*(4), 102250. https://doi.org/10.1016/j.ipm.2020.102250

Preeti, Kumar, M., & Sharma, H. K. (2023). A GAN-Based Model of Deepfake Detection in Social Media. *Procedia Computer Science*, *218*, 2153–2162. https://doi.org/10.1016/j.procs.2023.01.191

Rastogi, S., & Bansal, D. (2022). Disinformation detection on social media: An integrated approach. *Multimedia Tools and Applications*, *81*(28), 40675–40707. https://doi.org/10.1007/s11042-022-13129-y

Saurwein, F., & Spencer-Smith, C. (2020). Combating Disinformation on Social Media: Multilevel Governance and Distributed Accountability in Europe. *Digital Journalism*, *8*(6), 820–841. https://doi.org/10.1080/21670811.2020.1765401

Shankar, V. (2024). Managing the Twin Faces of AI: A Commentary on "Is AI Changing the World for Better or Worse?". *Journal of Macromarketing*, *44*(4), 892–899. https://doi.org/10.1177/02761467241286483

Shevlin, H., Vold, K., Crosby, M., & Halina, M. (2019). The limits of machine intelligence: Despite progress in machine intelligence, artificial general

intelligence is still a major challenge. *EMBO Reports*, *20*(10), e49177. https://doi.org/10.15252/embr.201949177

Shu, K., Bhattacharjee, Amrita, Alatawi, Faisal, Nazer, Tahora, Ding, Kaize, Karami, Mansooreh, & Liu, Huan. (2020). Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(6), e1385. https://doi.org/10.1002/widm.1385

Simchon, A., Sutton, A., Edwards, M., & Lewandowsky, S. (2023). Online reading habits can reveal personality traits: Towards detecting psychological microtargeting. *PNAS Nexus*, *2*(6), 1–9. https://doi.org/10.1093/pnasnexus/pgad191

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, *22*(11), 1531–1543. https://doi.org/10.1080/1369118X.2019.1648536

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised Fake News Detection on Social Media: A Generative Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33015644