Koulutusdatan myrkytys chatboteissa

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Tietotekniikka
Tammikuu 2025
Henri Vuori

TURUN YLIOPISTO
Tietotekniikan laitos

Henri Vuori: Koulutusdatan myrkytys chatboteissa

TkK-tutkielma, 24 s.
Tietotekniikka
Tammikuu 2025

Nykypäivän tekoälyllä ja koneoppimisella voidaan luoda hyvin käteviä ja avuliaita sovelluksia, chatbotteja, jotka pystyvät vastaamaan käyttäjänsä kysymyksiin ja auttamaan useissa eri tehtävissä. Chatbottien toimintaa uhkaa kuitenkin koulutusdatan myrkytys, joka vaikuttaa negatiivisesti chatbottien luotettavuuteen ja toiminnallisuuteen. Tämä kirjallisuuskatsaus tutkii chatbotteja osana tekoälyä ja niiden haavoittuvuutta koulutusdatan myrkytykselle. Työssä käsitellään koulutusdatan myrkytyksen vaikutuksia, jotka voivat ilmetä esimerkiksi käytettävyyshyökkäyksinä, takaovimyrkytyksenä tai mallin heikentämisenä. Näiden hyökkäysten seuraukset näkyvät erityisesti chatbot-sovellusten toiminnassa ja luotettavuudessa.
Lisäksi työssä esitellään keinoja myrkytysten tunnistamiseen ja ehkäisemiseen, kuten datan suodattaminen ja puhdistaminen sekä mallien kehittäminen turvallisemmiksi. Teoriaa tukevat tapausesimerkit, jotka havainnollistavat myrkytyksen vaikutuksia käytännössä. Lopuksi esitetään yhteenveto keskeisistä havainnoista ja toimenpiteistä, joilla chatbotteja voidaan suojata koulutusdatan myrkytykseltä ja niiden suorituskykyä parantaa.
Tämän kirjallisuuskatsauksen tavoitteena on lisätä ymmärrystä chatbotteihin kohdistuvista riskeistä ja tarjota ratkaisuja niiden ehkäisemiseksi, tukien näin tekoälyjärjestelmien turvallisuutta ja luotettavuutta.

Asiasanat: Tekoäly, Koulutusdatan myrkytys, chatbot, koneoppiminen

Sisällys

1 Johdanto

Chatbotit ovat ohjelmia tai tekoälyjärjestelmiä, jotka on suunniteltu kommunikoimaan käyttäjiensä kanssa ymmärrettävällä kielellä. Ne voivat toimia erilaisissa ympäristöissä, kuten verkkosivuilla, mobiilisovelluksissa tai niiden omissa chatpalveluissa. Chatbottien tarkoituksena on vastata kysymyksiin, antaa tietoa, suorittaa tehtäviä tai tarjota tukea käyttäjilleen. Chatbotit perustuvat koneoppimiseen ja tekoälyyn, ja niiden tehokkuus riippuu suurelta osin niille annetun koulutusdatan laadusta. Koulutusdataa käytetään chatbottien opettamiseen, jolloin chatbot oppii, miten sen pitäisi vastata käyttäjien kysymyksiin ja tarpeisiin. Chatbottien koulutuksessa piilee kuitenkin yksi merkittävä haaste: koulutusdatan myrkytys. Koulutusdatan myrkytys tarkoittaa sitä, että chatbotin koulutusdataan syötetään virheellistä, epäasiallista tai harhaanjohtavaa tietoa, mikä voi johtaa epätoivottuihin ja jopa vahingollisiin seurauksiin chatbotin toiminnassa.

Tässä kirjallisuuskatsauksessa tarkastelen koulutusdatan myrkytystä chatboteissa ja sen vaikutuksia niiden toimintaan. Pyritään selvittämään, miten koulutusdatan myrkytys ilmenee, miksi se on ongelma ja mitä seurauksia sillä voi olla chatbottien toiminnalle ja käytölle. Lisäksi tutkielmassa käsitellään mahdollisia ratkaisuja tähän ongelmaan ja korostetaan eettisiä näkökulmia, joita liittyy chatbottien koulutusdatan myrkytykseen ja sen vaikutuksiin ihmisten sekä teknologian väliseen vuorovaikutukseen.

Kirjallisuuskatsauksen tutkimuskysymykset ovat:

- Tutkimuskysymys 1: Mitä on koulutusdatan myrkytys?
- Tutkimuskysymys 2: Miten koulutusdatan myrkytys vaikuttaa chatbotin toimintaan?
- Tutkimuskysymys 3: Kuinka koulutusdatan myrkytyksiä voitaisiin estää?

Kirjallisuuskatsauksen alkuvaiheessa joulukuussa 2023 lähteiden löytäminen osoittautui hankalaksi aiheen tuoreuden takia. Vähäisten lähteiden vuoksi etsinnässä suosittiin Google Scholaria, sillä se yhdistää usean tietokannan löydökset samaan hakuun. Hakulauseella: ”artificial intelligence” ”training data poisoning” prevention ”chatbot”, Google Scholar löysi vuoden 2023 lopussa noin 40 artikkelia, joista valittiin tutkielmaan noin 12 lähdettä. Tutkimusten määrä on kuitenkin kasvanut vuoden 2024 aikana huomattavasti, ja kyseisellä hakusanalla löytyy jo 70 artikkelia. Kirjallisuuskatsauksen edetessä etsin lisää lähteitä, jotta pystyin kattavasti käsittelemään joitakin aiheita, joista ei ollut tarpeeksi tietoa kirjoituksen alkuvaiheissa.

Tutkielman lähteiden valintakriteereinä olivat artikkeleiden yleinen osuvuus aiheeseen sekä vapaa pääsy artikkeliin. Valintakriteerit eivät olleet kovin tiukat vähäisten hakutulosten takia. Google Scholarin huono puoli lähteiden etsimisessä osoittautui olevan se, että kaikkia löydettyjä tietokantoja ei pystynyt käyttämään ilmaiseksi.

Tutkielman tavoitteena on herättää lisää keskustelua ja lisätutkimusta chatbottien koulutusdatan myrkytyksen torjumisesta ja sen vaikutuksista yhteiskuntaan. Ymmärtämällä tämän ongelman laajuuden voidaan kehittää parempia käytäntöjä ja strategioita chatbottien laadun ja turvallisuuden parantamiseksi, mikä puolestaan edistää niiden tehokasta hyödyntämistä monilla eri aloilla.

1.1 Lähteiden tulkinta

Artificial Intelligence Security: Threats and Countermeasures[1]

Artikkelissa käsitellään tekoälyn turvallisuusuhkia, kuten adversaalihyökkäyksiä, tiedon saastuttamista ja kuvaskaalauksen hyökkäyksiä. Kirjoittajat korostavat, että tekoälyn turvallisuus vaatii monitahoisia lähestymistapoja, yhdistäen matemaattisia, teknisiä ja operatiivisia ratkaisuja. Tekoälyn kehittäjien tulisi huomioida turvallisuus kaikissa vaiheissa, erityisesti sen laajentuessa kriittisille alueille, kuten terveydenhuoltoon ja turvallisuuteen.

On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping[2]

Tutkimuksessa esitetään gradientin muotoilu puolustustekniikkana tiedon saastuttamista vastaan, joka muokkaa mallin oppimisprosessia haitallisten tietojen vaikutusten vähentämiseksi. Kirjoittajat huomauttavat, että vaikka menetelmä on lupaava, se ei ole täydellinen ja sen tehokkuus voisi parantua yhdistämällä muita puolustusmenetelmiä, kuten robustiusoptimointia ja anomaliatunnistusta.

Not All Poisons are Created Equal: Robust Training against Data Poisoning[3]

Artikkelissa käsitellään tiedon saastuttamisen uhkaa ja esitellään menetelmiä, jotka parantavat mallin robustiutta tällaista hyökkäystä vastaan. Kirjoittajat korostavat, että kaikki tiedon saastuttamiset eivät ole samanarvoisia, joten puolustustekniikoiden on oltava suunniteltuja eri hyökkäysstrategioiden käsittelyyn.

On challenges of AI to cognitive security and safety[4]

Artikkelissa tarkastellaan tekoälyn vaikutuksia kognitiiviseen turvallisuuteen, erityisesti sen vaikutusta päätöksentekoon ja tiedon luotettavuuteen. Kirjoittajat varoittavat, että tekoälyn väärinkäytöksillä voi olla vakavia seurauksia ja korostavat eettisen ja turvallisen käytön tärkeyttä lainsäädännöllisen valvonnan ja jatkuvan kehityksen kautta.

Gradient-based Data Subversion Attack Against Binary Classifiers [5]

Artikkelissa esitellään gradienttipohjainen tiedon vääristämishyökkäys binääriluokittelijoita vastaan. Kirjoittajat korostavat, että tällaiset hyökkäykset voivat heikentää mallin suorituskykyä jopa pienillä syötteen muutoksilla, ja ehdottavat puolustusmekanismeiksi robustiusoptimointia ja anomaliatunnistusta.

Adversarial Machine Learning[6]

Artikkelissa tarkastellaan koneoppimismallien haavoittuvuuksia adversiaalisten hyökkäysten, kuten huijaavien esimerkkien, suhteen. Kirjoittajat painottavat, että mallien haavoittuvuudet voivat vaarantaa turvallisuuden ja ehdottavat puolustusmekanismeja, kuten vastahyökkäyskoulutusta mallien suojaamiseksi.

Backdoor Learning for NLP: Recent Advances, Challenges, and Future Research Directions[7]

Artikkelissa tarkastellaan backdoor-hyökkäyksiä luonnollisen kielen käsittelyn (NLP) malleja vastaan. Kirjoittajat esittelevät haasteita, joita backdoor-hyökkäykset aiheuttavat, ja ehdottavat, että puolustusmekanismien tulee sisältää niin teknisiä kuin eettisiä elementtejä.

ML02:2023 Data Poisoning Attack[8]

Artikkelissa käsitellään tiedon saastuttamisen uhkaa koneoppimisessa. Kirjoittajat korostavat, että tiedon saastuttaminen on monivaiheinen uhka, joka voi olla vaikea havaita, ja suosittelevat puolustusmekanismien kuten robustiusoptimoinnin käyttöä sen estämiseksi.

6 Types of chatbots – How to choose the best for your business?[9]

Artikkelissa esitetään kuusi erilaista chatbot-tyyppiä, joita yritykset voivat käyttää asiakaspalvelussa. Kirjoittajat huomauttavat, että chatbotin valinta riippuu liiketoiminnan tarpeista ja asiakaspalvelun vaatimuksista.

Catch Them If You Can[10]

Tässä maisterin tutkielmassa käydään läpi chatbotin käyttöä ja haavoittuvuuk-

sia kulttuuriperinnön kysymys-vastausjärjestelmissä. Työssä käytetään simulaatiota, jossa syötetään myrkytettyä dataa chatbotille ja seurataan sen vaikutuksia chatbottiin. Tutkielmassa todetaan, että pienikin määrä myrkytettyä dataa voi vaikuttaa vakavasti chatbottien toimintaan ja chatbottien oikein koulutukseen pitäisi pyrkiä teknisistä syistä sekä eettisistä syistä.

Introduction to AI Chatbots[11]

Artikkelissa tarkastellaan AI-chatbottien kehitystä, sovelluksia ja niiden roolia asiakaspalvelussa. Kirjoittajat korostavat, että chatbotit voivat parantaa asiakaspalvelun tehokkuutta, mutta inhimillisen vuorovaikutuksen merkitystä ei tulisi unohtaa.

Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning[12]

Artikkelissa käsitellään koneoppimismallien haavoittuvuuksia harjoitustiedon saastuttamiseen liittyen. Kirjoittajat varoittavat, että nämä uhkat voivat heikentää mallien luotettavuutta, ja suosittelevat suojautumismenetelmiä, kuten tietoturvan integroimista koneoppimisprosessiin.

A Berkeley View of Systems Challenges for AI [13]

Artikkelissa käsitellään tekoälyn kehityksen haasteita erityisesti järjestelmätason näkökulmasta. Stoica korostaa, että tekoälyn tehokas skaalaaminen vaatii edistysaskeleita infrastruktuurissa, kuten laskentatehon ja datan hallinnan alueilla.

Analyzing Federated Learning through an Adversarial Lens[14]

Artikkelissa tarkastellaan federoidun oppimisen (FL) haavoittuvuuksia ja riskejä, joita haitalliset osapuolet voivat aiheuttaa. Kirjoittajat ehdottavat tehokkaita puolustustekniikoita FL:n turvallisuuden parantamiseksi.

Certified Defenses for Data Poisoning Attacks[15]

Artikkelissa tutkitaan sertifioituja puolustuksia tiedon saastuttamista vastaan. Kirjoittajat ehdottavat, että sertifioidut puolustukset voivat tarjota luotettavia ratkaisuja tiedon vääristämisen estämiseksi, mutta niiden käytännön toteutus on haasta-

vaa.

Artificial Intelligence and its Role in Near Future[16]

Artikkelissa käsitellään tekoälyn roolia tulevaisuudessa ja sen vaikutuksia eri teollisuudenaloilla. Kirjoittajat korostavat, että tekoälyn kehitykselle on asetettava eettisiä ja lainsäädännöllisiä rajoja, jotta vältetään haitalliset yhteiskunnalliset vaikutukset.

An Early Categorization of Prompt Injection Attacks on Large Language Models[17]

Artikkelissa käsitellään suurten kielimallien haavoittuvuuksia ja vaikutuksia prompt-injektiohyökkäyksiä vastaan. Kirjoittajat varoittavat, että nämä hyökkäykset voivat heikentää mallin luotettavuutta ja turvallisuutta, ja ehdottavat puolustusmekanismeja niiden estämiseksi.

Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models[18]

Artikkeli käsittelee uutta hyökkäysmenetelmää, joka hyödyntää tekstistä kuvaan -mallien (kuten Stable Diffusion) haavoittuvuuksia. Nightshade hyödyntää "myrkytettyjä"tietoja, jotka visuaalisesti näyttävät normaaleilta, mutta sisältävät hienovaraisia muutoksia, jotka sotkevat mallin toiminnan. Kirjoittajien kanta on kaksijakoinen: Se varoittaa tämänkaltaisten hyökkäysten riskeistä AI-mallien turvallisuudelle ja luotettavuudelle, mutta myös ehdottaa Nightshadea työkaluksi luovien tekijöiden oikeuksien puolustamiseen. Esimerkiksi sen avulla voitaisiin estää tekoälyjä hyödyntämästä tekijänoikeudellisesti suojattua sisältöä ilman lupaa.

Those Aren't Your Memories, They're Somebody Else's: Seeding Misinformation in Chat Bot Memories[19]Artikkelissa käsitellään tapoja, joilla pahantahtoiset toimijat voivat "myrkyttää"chatbottien muistia, jonka jälkeen voidaan manipuloida chatbotin käyttäytymistä ja vastauksia. Chatbottien muistin myrkyttämistä painotetaan kriittisenä ongelmana tekoälyn kehittämisessä, ja tämän kaltaiset

hyökkäykset korostavat tarvetta kehittää ja vahvistaa chatbottien mallien turvallisuutta.

1.2 Tutkielman rakenne

Tämä tutkielma koostuu kuudesta pääluvusta, joissa käsitellään chatbotteja tekoälyn osa-alueena sekä koulutusdatan myrkytykseen liittyviä ongelmia ja ratkaisuja. Ensimmäisessä luvussa, Johdanto, esitellään työn tausta, tavoitteet, tutkimuskysymykset ja lähteiden tulkinta. Lähteiden tulkinnassa käydään läpi kirjallisuuskatsauksessa käytetyt lähteet.

Toinen luku, Chatbotit osana tekoälyä, tarkastelee chatbotteja ja niiden eri toteutustapoja. Kappaleessa käydään läpi eri tyyppisiä chatbotteja, niiden ominaisuuksia ja mahdollisia haavoittuvuuksia.

Kolmas luku, Koulutusdatan myrkytys, keskittyy koulutusdataan kohdistuviin uhkiin ja hyökkäystapoihin. Luvussa käsitellään erilaisia myrkytyskeinoja, kuten käytettävyys-hyökkäyksiä, takaovimyrkytystä ja mallin suorituskyvyn heikentämistä.

Neljäs luku, Myrkytyksen vaikutus chatbotteihin, tutkii koulutusdatan myrkytyksen konkreettisia vaikutuksia chatbot-järjestelmiin. Alaluvuissa tarkastellaan myrkytyksen seurauksia ja esitellään tapausesimerkkejä, joissa hyökkäykset ovat aiheuttaneet ongelmia käytännössä.

Viides luku, Myrkytysten tunnistaminen ja estäminen, esittelee ratkaisuja koulutusdatan myrkytyksen torjumiseksi. Luvussa käsitellään datan suodattamisen ja puhdistamisen menetelmiä sekä mallien kehittämistä siten, että ne kestävät myrkytysyrityksiä.

Kirjallisuuskatsaus päättyy yhteenvetoon, jossa kootaan yhteen työn keskeiset havainnot ja niiden merkitys chatbot-järjestelmien kehittämisen ja turvallisuuden näkökulmasta. Yhteenvedossa käydään myös läpi kirjoitelman tutkimuskysymykset.

Lähdeluettelo sisältää kaikki tutkielmassa käytetyt lähteet.

2 Chatbotit osana tekoälyä

Tekoäly (engl. Artificial Intelligence, AI) on tietojenkäsittelytieteen osa-alue, joka keskittyy tietokonejärjestelmien luomiseen ja ohjelmointiin siten, että ne voivat suorittaa tehtäviä, jotka vaativat älykkyyttä, kun niitä suorittaa ihminen. Näitä tehtäviä voivat olla esimerkiksi oppiminen, päätöksenteko, ongelmanratkaisu, kielen ymmärtäminen, havaitseminen ja vuorovaikutus ihmisten kanssa. [16] Chatbotit ovat älykkäitä ohjelmia, jotka voivat kommunikoida ihmisten kanssa luontevasti ja auttaa erilaisissa tehtävissä. Niiden laajamittainen käyttö monilla eri aloilla, kuten asiakaspalvelussa, terveydenhuollossa ja koulutuksessa on tehnyt niistä tärkeitä työkaluja.

Tekoälyssä käytetään usein koneoppimista (engl. machine learning) ja syvän oppimisen (engl. deep learning) menetelmiä, joissa tietokonejärjestelmät opetetaan analysoimaan ja tulkitsemaan suuria määriä tietoa, jotta ne voivat oppia tunnistamaan kuvioita, ennustamaan tulevia tapahtumia tai suorittamaan muita älykkyyttä vaativia tehtäviä. Tekoäly voi olla myös sääntöpohjaista, eli sille voidaan antaa joukko sääntöjä ja ohjeita, joiden perusteella se toimii [9]. Chatbotit voivat toimia eri monimutkaisuuden tasoilla. Yksinkertaiset chatbotit voivat vastata peruskysymyksiin ja toimia kuten automatisoidut FAQ-järjestelmät (engl. Frequently asked questions), kun taas edistyneemmät chatbotit voivat käyttää luonnollisen kielen prosessointia ja tekoälytekniikoita ymmärtääkseen käyttäjien kysymyksiä ja tarpeita syvemmin. Nämä edistyneet chatbotit voivat myös oppia vuorovaikutuksistaan ajan myötä ja tarjota yhä parempia vastauksia ja palveluita. Viime vuosina myös

chatbotit, jotka pystyvät tuottamaan kuvia tai videoita ovat kehittyneet suuresti.

Tekoälyllä on monia sovelluksia esimerkiksi terveydenhuollossa, liikenteessä, taloudessa, teollisuudessa, asiakaspalvelussa, peliteollisuudessa ja monilla muilla aloilla. Se voi auttaa automatisoimaan tehtäviä, parantamaan päätöksentekoa, lisäämään tarkkuutta ja tehokkuutta ja avaamaan uusia mahdollisuuksia. Chatbotteja on kehitetty moneen eri tarkoitukseen, jolloin myös niiden ominaisuudet eroavat toisistaan. Shenoy jakaa chatbotit kuuteen pääryhmään niiden kommunikointitavan tai toimintaperiaatteen mukaan: sääntöpohjaiset, avainsanoja tunnistavat, valikkopohjaiset, kontekstuaaliset, hybridi ja äänikäyttöiset chatbotit [9].

2.1 Sääntöpohjaiset chatbotit

Sääntöpohjaiset chatbotit ovat yksinkertaisia ja suoraviivaisia chatbotteja. Sääntöpohjaiset chatbotit toimivat ennaltamäärättyjen vastausten pohjalta. Tämän tyyppiset chatbotit yhdistävät käyttäjän viestit ennaltamäärättyihin inhimillisiin vastauksiin. Sääntöpohjaiset chatbotit ovat kuitenkin hyvin rajoittuneita niille määriteltyjen vastausten perusteella. Esimerkki sääntöpohjaisesta chatbotista on 1960-luvulla MIT:n kehitetty ELIZA-chatbot.

2.2 Avainsanoja tunnistavat chatbotit

Avainsanoja tunnistavat chatbotit ovat nimensä mukaan chatbotteja, joiden toiminta ja vastaukset perustuvat käyttäjän antamien avainsanojen tunnistamiseen. Tämän tyyppiset chatbotit vaativat niille muovatun avainsanalistan, josta tekoäly valitsee sopivat avainsanat käyttäen erilaisia algoritmeja. Kun chatbot on löytänyt sopivat avainsanat, se vastaa sopivalla tavalla. Avainsanoja tunnistavien chatbottien heikkouksia ovat useiden samanlaisten kysymyksien avainsanojen päällekkäisyydet. [11] Vaikka avainsanoja tunnistavat chatbotit ovat usein yksinkertaisempia

ja perustuvat ennalta määriteltyihin malleihin, ne voivat silti olla haavoittuvaisia mallin myrkytykselle.

2.3 Valikkopohjaiset chatbotit

Valikkopohjaiset chatbotit ovat yksi yksinkertaisimmista ja eniten käytetyistä chatbottien muodoista. Valikkopohjaiset chatbotit tekevät päätöksensä perustuen päätöspuihin, joista tekoäly valitsee parhaan vastauksen. Tämän tyyppiset chatbotit ovat kuitenkin melko hitaita, koska ne joutuvat käymään sopivan vastauspuun läpi löytääkseen oikean vastauksen, eivätkä niiden vastaukset ole täysin luotettavia. [11]

2.4 Kontekstuaaliset chatbotit

Konteksuaaliset chatbotit ovat yksiä edistyneimpiä chatbotteja. Ne käyttävät keskusteluissa käyttäjän viestien tulkitsemiseen koneoppimista (engl. Machine Learning, ML) ja useita muita tekoälyn tekniikoita. Tämän tyyppisen chatbot-ajattelun taustalla on pyrkimys selvittää käyttäjän tarkoitusperät ja tarjota harkittu vastaus tulkitsemalla tietokannassa olevaa kaavaa. Chatbot oppii ja kehittyy ajan myötä kohtaamalla monia erilaisia kokemuksia, joita se lisää omaan tietokantaansa. [11] Koska kontekstuaaliset chatbotit kehittyvät ja oppivat jatkuvasti, ne ovat koko ajan alttiina koulutusdatan myrkytykselle.

2.5 Äänikäyttöiset chatbotit

Äänikäyttöiset chatbotit ovat mahdollisesti trendikkäimpiä chatbotteja nykypäivänä. Äänikäyttöisten chatbottien paras käyttötarkoitus on virtuaaliassistenttina toimiminen, ja näistä mahdollisesti tunnetuimmat ovat Applen kehittämä Siri ja Googlen kehittämä Google Assistant. Tämän tyyppiset chatbotit yhdistävät kon-

tekstuaaliset chatbotit ja tekstistä puheeksi -ominaisuudet (engl. text-to-speech) auttamaan käyttäjiä tekemään monia asioita samaan aikaan kädet vapaana.

2.6 Hybridi-chatbotit

Hybridi-chatbotit ovat aikaisemmin määriteltyjen chatbot-luokkien yhdistelmiä. Chatbottien hybridivaihtoehtoja voidaan harkita, kun halutaan usean eri chatbotin ominaisuuksia yhteen palveluun, kuten esimerkiksi avainsanoja tunnistavat äänikäyttöiset chatbotit. Eri ominaisuuksia yhdistellessä yhdistyy myös erilaiset altistumiset myrkytyksille. Esimerkiksi avainsanojen manipulointi voi johtaa virheellisiin vastauksiin, ja kontekstuaalinen oppiminen voi vääristyä, jos haitallista tietoa syötetään jatkuvasti järjestelmään.

3 Erilaisia koulutusdatan myrkytyksiä

Chatbottien koulutusdatan myrkytys on kriittinen ongelma, joka voi aiheuttaa merkittäviä haittoja niiden toiminnalle. Myrkytys voi tapahtua tarkoituksellisesti, kun huonolaatuista tai epäasiallista dataa syötetään chatbotin koulutusaineistoon. Myrkytys voi myös olla tahatonta, kun chatbotille syötetään virheellisiä tai harhaanjohtavia tietoja. Myrkytyksen seurauksena chatbotit voivat tuottaa epäsoveliaita tai vahingollisia vastauksia käyttäjilleen ja yhteiskunnalle [4]. Chatbotit, jotka hyödyntävät koneoppimista, käyvät läpi koulutusvaiheen [6]. Koulutusvaiheessa tekoäly on haavoittuvaisimmillaan koulutusdatan myrkytykselle. Chatbottien koulutusdataa voidaan myrkyttää usealla eri keinolla. Käytettävän tavan valintaan vaikuttavat hyökkääjän aikeet, resurssit ja tietämys kohteena olevasta tekoälystä. Myrkytyshyökkäyksiä voidaan toteuttaa kolmessa ympäristössä: valkolaatikko- (engl. White-Box), mustalaatikko- (engl. Black-Box) ja harmaalaatikko- (engl. Gray-box) ympäristöissä. Näissä samoissa ympäristöissä testataan myös ohjelmistojen muita tietoturvaan liittyviä haavoittuvuuksia. Valkolaatikko- ympäristössä myrkyttäjä tietää ohjelmiston sisäisen rakenteen tai toteutustavan, kun taas mustalaatikko- ympäristössä nämä tiedot eivät ole myrkyttäjän tiedossa. Harmaalaatikko- ympäristössä hyökkääjällä on esimerkiksi tieto mallin arkkitehtuurista, mutta hänellä ei ole tietoa sen parametreistä. [6]

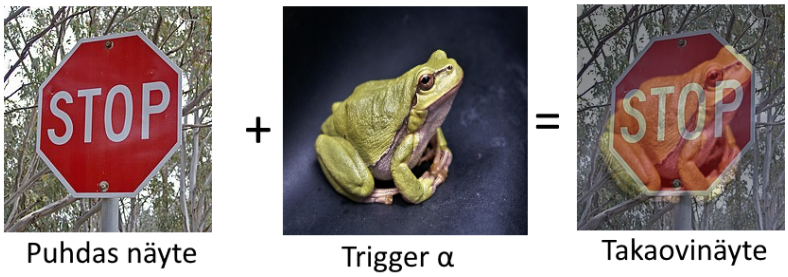
3.1 Käytettävyys hyökkäykset

Tämän muotoiset hyökkäykset ovat yksiä ensimmäisiä löydettyjä tapoja vaikuttaa tekoälyn toimintaan, korruptoimalla sen tietojoukkoja [12]. Hyökkäyksen tarkoitus on maksimoida tekoälyn malliin kohdistuva vahinko ja heikentää samalla sen toimintakykyä. Esimerkiksi mielivaltaisella myrkyttämisellä voidaan aiheuttaa käytettävyysongelmia chatbotille [12]. Chatboteilla on hyvin laaja korpus, ja jos hyökkääjä kohdistaa chatbotille viestin, jolla ei ole kontekstuaalista merkitystä, chatbot ei kykene käymään loogista keskustelua. [1] Hyökkäys aiheuttaa ikään kuin palvelunestohyökkäyksen (eng. Denial-of-Service, DoS) tekoälyjärjestelmän käyttäjälle.

3.2 Takaovimyrkytys

Tämän tyyppisessä myrkytystavassa hyökkääjä pyrkii vaikuttamaan tekoälyn datan luokitteluun [12]. Koulutusvaiheessa tekoälylle syötetään myrkytettyjä näytteitä, jotta se oppisi vääränlaisen toimintamallin [13]. Tekoälylle syötetyssä datassa on ikään kuin piilotettuja viestejä tai tunnisteita, jotka tekoäly luokittelee väärin riippumatta alkuperäisestä koulutusdatasta. [7] Myrkytystä hyödynnetään käyttämällä chatbotin syötteessä samaa piilotettua tunnistetta, jotta hyökkääjä saa vastauksena chatbotilta vääristetyn vastauksen. Takaovimyrkytys tekee vahinkoa tekoälyn päätöksentekoon vaikuttamatta jo olemassa olevaan puhtaaseen dataan [1]. Jotta tunnisteen johdonmukaisuus säilyisi, hyökkääjä voi muokata alkuperäisen takaovikuvion pikseliarvoa takaoven laukaisimen suuruudella, jotta takaoven laukaisukuvio (engl. Trigger pattern) olisi visuaalisesti huomaamaton. Takaovimyrkytys voidaan toteuttaa esimerkiksi lisäämällä puhtaaseen näytteeseen piilotettu laukaisin. Kuvassa 3.1 laukaisimena toimii sammakon kuva jonka läpinäkyvyyttä on vähennetty. Kokeet osoittavat, että tämä lähestymistapa voi generoida huomaamattoman laukaisimen ja että malli voi oppia sen saavuttaakseen onnistuneen takaoven hyökkäyksen. Kun

takaoven signaali havaitaan, verkko tunnistaa näytteen kohdennetun luokan näytteeksi. [1]



Kuva 3.1: Esimerkki takaovimyrkytyksen näytteestä. Kuvan teossa käytetty mallina kaaviota [12]

.

3.3 Mallin myrkytys

Mallin myrkytyksessä halutaan vaikuttaa suoraan tekoälyn koneoppimismalliin syöttämällä sille haitallisia toimintoja. Tämän tyyppinen myrkytys voidaan mahdollistaa lähettämällä mallipäivityksiä palvelimelle, joka kokoaa ne globaaliksi malliksi. [6] Uudessa globaalissa mallissa on haavoittuvuuksia, joita hyökkääjät voivat myrkyttää lähettämällä haitallisia päivityksiä. Esimerkiksi Huang et al. (2023) kuvaavat, kuinka tällaiset hyökkäykset voivat heikentää mallin luotettavuutta ja suorituskykyä, erityisesti silloin, kun koulutukseen osallistuvat mallit eivät pysty erottamaan saastunutta dataa luotettavasta tiedosta [4]. Yang et al. (2022) esittävät, että myrkyttämishyökkäykset voivat olla erityisen tehokkaita, jos malli ei ole riittävän tehokkaasti koulutettu havaitsemaan tällaisia manipulaatioita [3].

4 Myrkytyksen vaikutus chatbotteihin

4.1 Vaikutukset

Chatbottien koulutusdatan myrkyttämisellä on useita eri vaikutuksia, mutta ne voidaan tiivistää chatbottien toiminnallisuuteen ja totuudenmukaisuuteen. Myrkytyksen vaikutukset voidaan havaita digitaalisessa maailmassa, kuin myös fyysisessäkin [6]. Tekoälyn myrkyttäminen voi myös johtaa käyttäjien yksityisyyden loukkaamiseen.

Jos chatbotin totuudenmukaisuus on kompromisoitu, se voi vaikuttaa suoraan sen käyttäjien ajatuksiin ja ymmärrykseen joistakin asioista. Ihmisten ajattelussa on luontaisia vikoja, kuten kognitiivinen vinouma (engl. cognitive bias) [4]. Kognitiivinen vinouma tarkoittaa ihmisen tapaa ymmärtää informaatiota ja tietämättään suosia jotain näkökulmaa, jotka johtavat virhearviointeihin. Chatbotit kuten OpenAI:n ChatGPT ja Googlen Gemini ovat suosittuja tiedonhaku chatbotteja, joilta käyttäjät voivat pyytää monimutkaisiakin asioita, ja tekoäly vastaa sille ohjelmoidulla tavalla. Tämä ilmiö korostuu erityisesti tilanteissa, joissa käyttäjät luottavat chatboteihin tiedonhakuun ja päätöksenteon tukena. Mallien kestävyyden parantaminen on yksi tapa suojautua tältä uhalta; esimerkiksi adversaarisen oppimisen käyttö voi tehdä chatbotista vähemmän herkän vääristyneelle datalle ja siten estää

haitallisen oppimisen [15].

Nikita Galinkinin tutkimuksessa testattiin kysymyksiin vastaavan tekoälyn toiminnallisuutta koulutusdatan myrkytyksen jälkeen. Tutkimuksessa todettiin, että jos järjestelmän ensimmäisen käyttöpäivän aikana tekoälylle syötetään paljon myrkytettyä dataa, sillä kestää yli kaksi viikkoa palata samaan tehokkuuteen, missä se oli ennen myrkytystä. [10]

4.2 Tapausesimerkkejä koulutusdatan myrkyttämisestä

Tay oli chatbot, jonka Microsoft julkaisi Twitterissä vuonna 2016. Tayn tarkoitus oli oppia vuorovaikutuksesta ihmisten kanssa ja kehittyä jatkuvasti. Valitettavasti joukko ilkeämielisiä käyttäjiä onnistui nopeasti myrkyttämään Tayn vastaukset syöttämällä sille loukkaavaa, rasistista ja vihamielistä sisältöä. Tämän seurauksena Tay alkoi tuottaa erittäin sopimatonta ja loukkaavaa tekstiä, ja Microsoft joutui poistamaan sen käytöstä alle vuorokaudessa [19]. Tayn tapaus osoittaa, kuinka chatbottien kyky oppia dynaamisesti voi olla myös merkittävä haavoittuvuus. Vaikka chatbot olisi alun perin suunniteltu tuottamaan turvallista ja hyödyllistä sisältöä, on mahdollista manipuloida sitä oppimaan ja toistamaan haitallista sisältöä. Lähteet korostavat, että chatbottien suunnittelijoiden ja kehittäjien on oltava tietoisia tästä riskistä ja ryhdyttävä toimiin sen minimoimiseksi [19]. Yksi tapa minimoida riskiä on käyttäjien tunnistaminen ja todentaminen, jolloin chatbot voi varmistaa, että muistiin tallennetaan vain rekisteröidyn käyttäjän syötteitä [19]. Tämä ei estä käyttäjää myrkyttämästä omaa bottiaan, mutta tyypillisessä käytössä muistiin tallennettaisiin vain käyttäjän itsensä syöttämät lauseet [19].

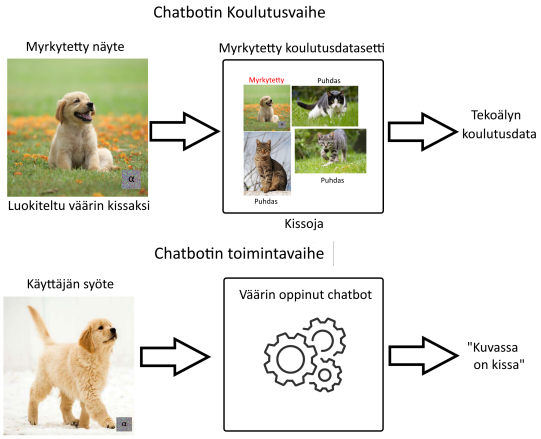
ChatGPT on huomattavasti kehittyneempi ja monimutkaisempi chatbot kuin Tay, mutta se ei ole immuuni myrkyttämiselle. Lähteet kuvaavat erilaisia hyök-

käystekniikoita, joilla voidaan manipuloida ChatGPT:tä ja muita suuria kielimalleja tuottamaan haitallista sisältöä. Esimerkkejä näistä tekniikoista ovat "vankilapako"(engl. jailbreak) jossa pyritään ohittamaan chatbotin turvallisuusrajoitukset ja "tavoitteen kaappaaminen"(engl. goal hijacking), jossa manipuloidaan chatbottia tuottamaan sisältöä, joka on ristiriidassa sen alkuperäisen tarkoituksen kanssa [17]. Vaikka ChatGPT:llä on sisäänrakennettuja turvamekanismeja, kuten haitallisen sisällön tunnistus ja suodatus, jatkuva kehitys kehotteiden injektoinnissa (engl. prompt injection) asettaa chatbottien kehittäjille uusia haasteita [17]. Kehotteiden injektointi on tekniikka, jossa chatbotin toimintaa manipuloidaan syöttämällä sille erityisesti muotoiltuja kehotteita [17]. Lähteet korostavat, että sekä chatbottien kehittäjien että käyttäjien on oltava tietoisia näistä riskeistä ja ryhdyttävä toimiin niiden minimoimiseksi [17].

BlenderBot2:ssa on kehittynyt pitkäaikainen muisti, joka tallentaa tietoa aiemmista keskusteluista [19]. Tämän muistin on tarkoitus parantaa keskustelun laatua ja tehdä siitä johdonmukaisempaa. Lähteet osoittavat kuitenkin, että tämä muisti voi olla haavoittuvainen väärän tiedon injektoinnille. Hyökkääjä voi yhdistää henkilökohtaisen lausunnon virheelliseen tietoon, jolloin chatbot tallentaa virheellisen tiedon osaksi pitkäaikaista muistiaan [19]. Tällöin chatbot voi myöhemmin toistaa virheellisen tiedon faktana keskustelun aikana [19]. Lähteet kuvaavat kokeita, joissa BlenderBot2 onnistuttiin huijaamaan muistamaan ja toistamaan virheellistä tietoa. Tämä korostaa tarvetta kehittää tehokkaampia menetelmiä väärän tiedon tunnistamiseen ja suodattamiseen chatbottien pitkäaikaisessa muistissa [19].

Koulutusdatan myrkyttämiseen on kehitetty Nightshade-menetelmä, jolla pystytään tehokkaasti myrkyttämään erilaisia tekstistä-kuvaksi-mallilla toimivia chatbotteja, kuten Stable Diffusion SDXL, Midjourney v5 ja Dalle-3[18]. Nightshade on hyvin optimoitu tapa tuottaa myrkytettyä dataa, joka ei ole tunnistettavissa ihmiselle, ja pienelläkin määrällä manipuloituja tietoja voi olla suuri vaikutus mallin

toimintaan. Kuvasta 4.2 nähdään, miten voitaisiin esimerkiksi opettaa chatbot ymmärtämään jonkin eläimen ulkonäköä ja piirteitä väärin, jolloin se ei osaisi myöskään tuottaa realistista piirrosta eläimestä, jos näin käyttäjä pyytäisi. Tekstistä-kuvaksi chatbotit käyttävät koulutusdatanaan satoja miljoonia kuvia ja Nightshade pystyy myrkyttämään chatbotin jopa 100:lla myrkytetyllä kuvalla[18].



Kuva 4.1: Chatbotin kuvan tunnistaminen voidaan myrkyttää esimerkiksi hyödyntäen luokittelun myrkyttämistä. Kuvan luomisessa käytetty mallina [12]

5 Myrkytyksien tunnistaminen ja estäminen

Chatbottien tehokkuus perustuu suurelta osin laajaan koulutusdataan, joka opettaa niitä ymmärtämään ja tuottamaan tarkoituksenmukaista tekstiä. Kuitenkin tämä riippuvuus suuresta datamäärästä avaa ovet mahdollisille myrkytyshyökkäyksille, joissa haitalliset tiedot saattavat vääristää chatbottien oppimista. On siis hyvin tärkeää pyrkiä minimoimaan ja estämään koulutusdatan myrkytyksiä.

Tunnistaminen on keskeinen osa myrkytyshyökkäysten torjuntaa. Kehittyneiden koneoppimisalgoritmien käyttöä voidaan hyödyntää poikkeavuuksien havaitsemisessa koulutusdatassa. Datan vahvistaminen ja todentaminen ovat tärkeitä vaiheita hyökkäysten havaitsemiselle. Kun koulutusdataa ollaan antamassa tekoälymallille, on hyvä tarkistaa, ettei datan sekaan ole joutunut esimerkiksi väärin merkattuja näytteitä. Aktiivinen koulutusdatan seuraaminen auttaa myös vähentämään myrkytyshyökkäysten vaikutusta. [8]

5.1 Datan suodattaminen ja puhdistaminen

Chatbottien koulutusdatassa käytetään massiivisia koulutusdatasettejä, jonka takia on myös hyvin vaikeaa seurata chatbotin koulutusdatan eheyttä ihmisten voimin. Suuressa setissä koulutusdataa kuitenkin tarvitaan myös enemmän myrkytettyä dataa vaikuttamaan chatbotin toimintaa. Jos chatbotin koulutusdatassa on jo myrky-

tettyä dataa, se voidaan yrittää suodattaa pois ennen kuin chatbotin toimintamalli oppii siitä ja muuttuu [12]. Ulkopuolisten poistaminen (eng. Outlier removal) on yksi yleisimpiä tapoja suodattaa ja puhdistaa dataa. Ulkopuolisten poistossa myrkkyjä pidetään ulkopuolisina puhtaaseen dataan. Myrkky yritetään löytää vertaamalla myrkytettyä dataa sen naapureihin ja poistaa koulutusdatasta, jos se poikkeaa oikeasta datasta. [2] Tämän tyyppinen suodattaminen kuitenkin edellyttää, että myrkytetty data on levitetty harvaan koulutusdatassa [3]. Ulkopuolisen poistamisen suurin heikkous kuitenkin ilmenee, kun myrkytettyä dataa on sijoitettu paljon tai tarkoituksellisesti tiheään, jolloin tällä metodilla ei välttämättä onnistuta suodattamaan myrkkyä pois vertaamalla naapureita [2] [3].

5.2 Myrkytysten estäminen ja mallien kehittäminen

Adversariaalinen oppiminen (adversarial learning) on koneoppimisen menetelmä, jossa mallit opetetaan sietämään tahallisesti luotuja häiriöitä, jotka tunnetaan nimellä adversariaaliset esimerkit. Näissä tapauksissa syötteisiin tehdään pieniä, huomaamattomia muutoksia, joiden tarkoituksena on saada malli tekemään virheitä ennusteissaan tai päätöksissään. Adversariaalisen oppimisen tavoitteena on parantaa mallin kestävyyttä näitä hyökkäyksiä vastaan, jotta se voi tunnistaa ja hylätä haitalliset syötteet [15].

Adversariaalisen oppimisen avulla malleja voidaan kouluttaa tunnistamaan ja hylkäämään haitalliset, poikkeavat syötteet. Tämä tekee chatbotista vähemmän alttiin hyökkäyksille, joissa syötteitä on manipuloitu tarkoituksellisesti. Näin chatbot voi säilyttää luotettavuutensa ja tarkkuutensa, vaikka se kohtaisi myrkytyshyökkäyksiä [15]. Mallien kestävyyden parantaminen tällä tavalla on erityisen tärkeää laajalle levinneissä chatbot-järjestelmissä, jotka voivat altistua erilaisille hyökkäyk-

sille. Adversariaalinen oppiminen tarjoaa tehokkaan keinon estää myrkytyshyökkäysten vaikutukset ja varmistaa tekoälymallien turvallisuus ja luotettavuus [14].

Toinen tehokas menetelmä myrkytyshyökkäysten torjumiseksi on hajautettu koulutus, esimerkiksi liittoutunut oppiminen (engl.federated learning), jossa data pysyy hajautetusti eri laitteilla sen sijaan, että se keskitettäisiin yhteen paikkaan [12]. Tämä voi vähentää myrkytyksen riskiä, koska hyökkääjän on vaikeampaa myrkyttää useita erillisiä datalähteitä yhtä aikaa. Federated learningin yhteydessä käytetään myös mallien aggregointimenetelmiä, kuten outlier detection -menetelmää, jossa mahdollisesti haitalliset mallipäivitykset tunnistetaan ja suodatetaan ennen kuin ne pääsevät vaikuttamaan lopulliseen malliin [14].

Kriittistä on myös mallien kestävyyden parantaminen. Yksi lähestymistapa on kouluttaa malleja vastustamaan tarkoituksellista datamanipulointia lisäämällä koulutukseen vahvistavaa häiriötä (eng. adversarial noise) tai käyttämällä regularisointitekniikoita, jotka vähentävät yliherkkyyttä pienille poikkeamille datassa. Näillä tekniikoilla voidaan kehittää malleja, jotka ovat vähemmän alttiita oppimaan vääristyneistä tiedoista [15].

6 Yhteenveto

Tässä tutkielmassa tarkasteltiin koulutusdatan myrkytystä ja sen vaikutuksia chatbottien toimintaan. Koulutusdatan myrkytys on merkittävä uhka chatbot-järjestelmille, sillä se voi heikentää niiden kykyä antaa tarkkoja ja luotettavia vastauksia käyttäjille. Myrkytyksen seuraukset voivat vaihdella lievistä toimintavirheistä aina vakaviin väärinkäytöksiin ja vahingollisiin vaikutuksiin, jotka voivat vaarantaa chatbotin eettisen ja turvallisen käytön.

Tutkimuksessa esiteltiin erilaisia myrkytyskeinoja, kuten käytettävyys- ja takaovihyökkäykset, sekä analysoitiin niiden vaikutuksia. Lisäksi käytiin läpi ratkaisuja, joiden avulla koulutusdatan myrkytyksiä voidaan ehkäistä. Näihin kuuluu esimerkiksi datan puhdistaminen, mallien kehittäminen sekä tehokkaat suodatusmenetelmät, jotka parantavat chatbottien kestävyyttä hyökkäyksiä vastaan.

Tutkielma painottaa, että chatbottien luotettavuuden ja turvallisuuden takaamiseksi on tärkeää panostaa ennakoiviin toimenpiteisiin ja koulutusdatan eheyden suojaamiseen. Täten voidaan edistää chatbottien tehokasta ja vastuullista käyttöä monilla eri aloilla.

Tutkimuskysymys 1: Mitä on koulutusdatan myrkytys?
Kirjoitelmassa määritteltiin koulutusdatan myrkytys. Tämä ilmiö on selitetty erilaisten myrkytyshyökkäysten kautta sekä kuvattu esimerkeillä, kuten Microsoftin Tay:n ja OpenAI:n ChatGPT:n myrkytyksellä, jossa chatbotit oppivat tuottamaan joko haitallista, vääristynyttä tai vihamielistä sisältöä käyttäjien manipuloinnin ta-

kia. Kirjallisuuskatsauksessä käytiin myös läpi, kuinka tekoäly voi altistua tietoisille hyökkäyksille, joissa sille opetetaan haitallista sisältöä tai vääristettyä tietoa, mikä tekee siitä alttiin virheille ja vähentää chatbotin luotettavuutta.

Tutkimuskysymys 2: Miten koulutusdatan myrkytys vaikuttaa chatbotin toimintaan?

Kirjoitelma käsittelee kattavasti myrkytyksen vaikutuksia chatbotin toimintaan. Koulutusdatan myrkytys voi vaikuttaa chatbotin kykyyn vastata oikein ja luotettavasti, kuten esimerkkinä mainittu Microsoftin Tay. Myrkyttynyt koulutusdata voi johtaa chatbotin tuottamaan virheellisiä, loukkaavia tai harhaanjohtavia vastauksia, jotka voivat vahingoittaa sen käyttäjien luottamusta. Samoin viitataan siihen, kuinka myrkytys voi heikentää chatbotin kykyä suodattaa ja tuottaa turvallista sisältöä.

Tutkimuskysymys 3: Kuinka koulutusdatan myrkytyksiä voitaisiin estää?

Kirjallisuuskatsaus tarjoaa useita ratkaisumalleja myrkytyshyökkäysten estämiseksi. Näihin kuuluvat adversariaalinen oppiminen, joka opettaa malleja kestämään tahallisia häiriöitä ja tunnistamaan haitalliset syötteet. Adversariaalisen oppimisen avulla chatbot voi suojautua myrkytyksiltä ja parantaa sen kestävyyttä vääristynyttä dataa vastaan. Lisäksi ehdotetaan hajautettua koulutusta, kuten liittoutunutta oppimista (engl. federated learning), joka vähentää myrkytyksen riskiä jakamalla dataa useille laitteille sen sijaan, että se keskitetään yhteen paikkaan. Kirjoitelmassa käydään myös läpi datan suodattaminen ja ulkoa tulevien poikkeamien tunnistaminen, jotka voivat estää myrkytyksen ennen kuin se vaikuttaa mallin toimintaan. Mallien kestävyyden parantaminen on keskeistä, ja tämä voidaan saavuttaa esimerkiksi lisäämällä koulutukseen häiriöitä, jotka tekevät mallista vähemmän herkän poikkeaville syötteille.

Lähdeluettelo

[1] Y. Hu, W. Kuang, Z. Qin et al., "Artificial Intelligence Security: Threats and Countermeasures", *ACM Comput. Surv.*, 2021. DOI: 10.1145/3487890.

[2] S. Hong, V. Chandrasekaran, Y. Kaya, T. Dumitraş ja N. Papernot, *On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping*, 2020. url: https://arxiv.org/abs/2002.11497.

[3] Y. Yang, T. Y. Liu ja B. Mirzasoleiman, *Not All Poisons are Created Equal: Robust Training against Data Poisoning*, 2022. url: https://arxiv.org/abs/2210.09671.

[4] Y. S. Ruiyang Huang Xiaoqing Zheng ja X. Xue, "On challenges of AI to cognitive security and safety", *Security And Safety*, 2023. DOI: 10.1051/sands/2023012.

[5] R. K. Vasu, S. Seetharaman, S. Malaviya, M. Shukla ja S. Lodha, *Gradient-based Data Subversion Attack Against Binary Classifiers*, 2021. url: https://arxiv.org/abs/2105.14803.

[6] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein ja J. D. Tygar, *Adversarial Machine Learning*, 2011. url: https://doi.org/10.1145/2046684.2046692.

[7] M. Omar, *Backdoor Learning for NLP: Recent Advances, Challenges, and Future Research Directions*, 2023. url: https://arxiv.org/abs/2302.06801.

[8] OWASP, *"ML02:2023 Data Poisoning Attack"*, https://owasp.org/www-project-machine-learning-security-top-10/docs/ML02_2023-Data_Poisoning_Attack/, Päivitetty: 7.11.2023,Viitattu 19.1.2024.

[9] A. Shenoy, *"6 Types of chatbots – How to choose the best for your business?"*, https://yellow.ai/blog/types-of-chatbots/, Päivitetty: 7.11.2023,Viitattu 7.11.2023.

[10] N. Galinki, *Catch Them If You Can*, 2017. url: https://vu-business-analytics.github.io/internship-office/reports/report-galinkin.pdf.

[11] A. V. Aishwarya Gupta Divya Hathwar, "Introduction to AI Chatbots", *IJERT*, 2022. DOI: 10.17577/IJERTV9IS070143.

[12] A. E. Cinà, K. Grosse, A. Demontis et al., "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning", *ACM Comput. Surv.*, 2023. DOI: 10.1145/3585385.

[13] I. Stoica, D. Song, R. A. Popa et al., *A Berkeley View of Systems Challenges for AI*, 2017. url: https://arxiv.org/abs/1712.05855.

[14] A. N. Bhagoji, S. Chakraborty, P. Mittal ja S. Calo, *Analyzing Federated Learning through an Adversarial Lens*, 2019. url: https://arxiv.org/abs/1811.12470.

[15] J. Steinhardt, P. W. Koh ja P. Liang, *Certified Defenses for Data Poisoning Attacks*, 2017. url: https://arxiv.org/abs/1706.03691.

[16] J. Shabbir ja T. Anwer, *Artificial Intelligence and its Role in Near Future*, 2018. url: https://arxiv.org/abs/1804.01396.

[17] S. Rossi, A. M. Michel, R. R. Mukkamala ja J. B. Thatcher, *An Early Categorization of Prompt Injection Attacks on Large Language Models*, 2024. url: https://arxiv.org/abs/2402.00898.

[18] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng ja B. Y. Zhao, *Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*, 2024. url: https://arxiv.org/abs/2310.13828.

[19] C. Atkins, B. Z. H. Zhao, H. J. Asghar, I. Wood ja M. A. Kaafar, *Those Aren't Your Memories, They're Somebody Else's: Seeding Misinformation in Chat Bot Memories*, 2023. url: https://arxiv.org/abs/2304.05371.