Hei vaan kaikille olen Akseli Ahonen ja tänään pidän esitelmän tekoälyn myrkyttämisestä eli AI poisoningista. Tähän alkuun voisin sanoa, että Tää on tällä hetkellä erittäin ajankohtainen aihe, koska tekoäly on niin laajasti käytössä eri toimialoilla, kuten terveydenhuollossa, liikenteessä, robotiikassa, sosiaalisessa mediassa ja finanssialalla. Tän laajan käytettävyyden takia tekoäly on houkutteleva kohde kyberhyökkäyksille. Ja yksi vaarallisimmista uhista on tekoälyn myrkyttäminen

Mitä on tekoälyn myrkyttäminen?

Yhdellä lauseella ilmaistuna tekoälyn myrkyttäminen tarkoittaa.

- Tahallista tekoälyn toiminnan korruptointia

jossa tavoitteena on saada tekoäly tekemään virheellisiä tai haitallisia päätöksiä tai luomaan takaovia hyökkääjän edun mukaisesti.

Hyökkääjä voi esimerkiksi manipuloida kasvojentunnistusjärjestelmää siten, että se ei tunnista tiettyjä henkilöitä tai hyväksyy vääriä henkilöitä sisään.

- Kohdistuu dataan, heikkousiin algoritmissa tai itse malliin

Eli tekoälyn myrkyttäminen on prosessia, jossa haitallista dataa syötetään koneoppimismalliin. Tämä voi tapahtua koulutusvaiheessa tai toimintavaiheessa.

Algoritmien heikkoudet näkyvät sellaisissa tekoälyn malleissa, joka pitää käyttäjien datan erillään tietoturvasyistä ja kouluttaa useita malleja, jotka lopulta yhdistetään yhdeksi isommaksi kokonaisuudeksi. Tässä hyökkääjä voi luoda useita anonyymejä käyttäjiä ja vaikuttaa yhdistettyyn malliin. Käytännössä data aseistetaan tekoälyä vastaan.

Itse malliin kohdistuvat hyökkäykset toimivat siten, että valmiin mallin voi korvata myrkytetyllä mallilla käyttäen perinteisiä kyberhyökkäyksen menetelmiä.

-Tarvittava ”myrkytetyn” datan määrä on vähäistä, noin 1-3% koko datasta.

Toisin sanoen pienelläkin määrällä dataa voidaan huomattavasti alentaa tekoälyn tehokkuutta muodostaa tarkkoja tai oikeita lopputuloksia.

Menetelmiä

-Datan väärin luokittelua eli luokkien manipulointia

Hyökkääjä pyrkii vaihtamaan oikeat luokat vääriksi, jolloin malli oppii tunnistamaan luokat väärin ja lopputulos on väärä.

Esimerkiksi sähköpostin roskapostisuodatus käyttää tilastollista koneoppimista hyvän sähköpostin ja roskapostin erottamiseen. Muuttamalla vain 1 % koulutusviesteistä algoritmi voi tulla tehottomaksi. Hyökkääjät käyttävät termejä, jotka esiintyvät tavallisissa sähköposteissa, ja käyttävät niitä roskapostissa pakottaakseen uudelleenkoulutuksen aikana tapahtuvan luokittelun muuttamisen tuoreella tietojoukolla.

-Datan manipulointi

Datan manipulointi tarkoittaa tietojen muokkaamista koneoppimismallin koulutusjoukossa siten, että malli luokittelee tiedot virheellisesti tai käyttäytyy ennalta määrätyllä haitallisella tavalla vastauksena tiettyihin syötteisiin. Koulutustietojen manipuloimisen tekniikoita ovat muun muassa:

- Väärien tietojen lisääminen

- Oikeiden tietojen poistaminen

- Vihamielisten näytteiden lisääminen

- Vahingollisen datan syöttäminen

Voi olla esimerkiksi täysin väärää tietoa tai hieman muokattua tietoa, joka sisältää väärää tietoa. Voi myös olla vain mallin kannalta huonoa tietoa, joka johtaa ei toivottuun toimintaan.

Esittelemällä haitallisia tietoja hyökkääjät voivat estää tekoälyn toimimasta oikein. Vuonna 2016 Microsoft esitteli Tay:n, Twitter-botin, jonka tarkoituksena oli keskustella 18–24-vuotiaiden käyttäjien kanssa ”rentojen ja leikkisien keskustelujen” kautta. Tay oli ”oleellisesti robottipapukaija, jolla oli internet-yhteys” (Vincent, 2016). Ihmiset alkoivat twiitata botille sopimattomia kommentteja, ja 24 tunnin kuluttua sen julkaisusta Tay muuttui kiroilevaksi rasistiksi, antifeministiksi ja holokaustin kieltäjäksi. Microsoftin mukaan Tay rakennettiin käyttäen relevanttia, mallinnettua, puhdistettua ja suodatettua dataa. Kuitenkin, kun se otettiin käyttöön ja oppi julkisista tiedoista, sen käyttäytyminen muuttui radikaalisti.

Lopputuloksena on, että malli ei toimi halutusti.

Työntämällä korruptoitunutta tietoa hyökkääjät voivat manipuloida haittaohjelmien havaitsemisjärjestelmiä tunnistamaan hyvän tiedon haitalliseksi ja päinvastoin, sekä muokata sähköpostin roskapostisuodatusjärjestelmiä niin, että roskapostit kuten phishing-viestit pääsevät läpi, ja vääristää verkko-ohjelmasovelluksia siten, että valtuuttamaton tietokone saa luvan käyttää verkkoa.

- Laukaisimilla luodut takaovet

Tässä tilanteessa hyökkääjät piilottavat malliin erityisiä kaavoja tai ”laukaisijoita”. Kun nämä kaavat ilmenevät todellisessa käytössä, ne saavat mallin käyttäytymään odottamattomilla tavoilla. Kuvittele, että pysäytyskyljessä on tarra, joka hämmentää itseajavaa ajoneuvoa, jolloin se luulee sitä väistämismerkiksi eikä pysäytysmerkiksi.

Gu et al. (2019) osoittavat, että hyökkääjä voi luoda takaoven haavoittuvuuden liikennemerkki-luokittelijassa lisäämällä useita myrkytettyjä syötteitä koulutuksen aikana. Kun auto lähestyi pysäytysmerkkiä, jossa oli erityinen tarra, se tunnisti sen nopeusrajoitusmerkiksi.

Kyberturvallisuuden uhat

- Tekoälyä käytetään paljon muun muassa haittaohjelmien, roskapostin ja petoksen tunnistuksessa.

Sosiaalinen manipulointi: AI poisoning voi mahdollistaa väärän informaation levittämisen ja siten vaikuttaa yleiseen mielipiteeseen tai politiikkaan.

Tietovuodot: AI poisoning voi ohjata tekoälymallin paljastamaan luottamuksellista tai arkaluontoista tietoa väärissä tilanteissa.

Toiminnan sabotointi: Vahingoittunut tekoälymalli voi tahallisesti hidastaa, häiritä tai estää järjestelmien toimintaa, mikä voi lamauttaa esimerkiksi terveydenhuollon tai liikenteen järjestelmiä.

Suojautuminen

* Koulutuksessa käytetyn datan lähteiden hallinta

Pidetään huolta siitä, että data tulee luotettavista lähteistä ja varmistetaan että sitä dataa käytetään halutulla tavalla. Myös se, että hyökkääjä ei tiedä mistä tietoa kerätään auttaa pitämään datan puhtaana.

* Datan yksilöiminen vain yhdelle tekoälylle

Pyritään rajoittamaan datan käyttöä mahdolisimman pienelle määrälle tekoälymallille. Tähän syynä on se, että jos data korruptoituu ja sitä käytetään useaan eri mallin kouluttamiseen jokainen malleista, on korruptoitunut.

* Jatkuva toiminnan tarkkailu ja kirjanpito käytetystä datasta.

Tarkkaillan, että tekoälyn toiminta vastaa oletettua ja, jos huomataan vääränlaista toimintaa siihen pystytään reagoimaan.