Hei vaan kaikille olen Akseli Ahonen ja tänään pidän esitelmän tekoälyn myrkyttämisestä eli AI poisoningista. Kuten kaikki varmaan tietävät tekoäly on laajasti käytössä eri toimialoilla, kuten terveydenhuollossa, liikenteessä, robotiikassa, sosiaalisessa mediassa ja finanssialalla. Juuri tämän laajan käytettävyyden takia tekoäly on myös houkutteleva kohde kyberhyökkäyksille. Ja yksi vaarallisimmista uhista on tekoälyn myrkyttäminen

Mitä on tekoälyn myrkyttäminen?

Pähkinän kuoressa tekoälyn myrkyttäminen tarkoittaa.

- Tahallista tekoälyn toiminnan korruptointia

jossa tavoitteena on saada tekoäly tekemään virheellisiä tai haitallisia päätöksiä tai luomaan takaoven 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 AI poisoning tarkoittaa 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 lopputuloksia.

Menetelmiä

-Datan väärin luokittelua.

Hyökkääjä pyrkii vaihtamaan oikeat luokat vääriksi, jolloin malli oppii tunnistamaan luokat väärin ja lopputulos on väärä. Yleinen esimerkki on Nightshade työkalu, jolla artistit voivat muokata kuviensa pixeleitä hieman, jolloin jos niitä käytetään tekoälyn kouluttamisessa niin lopputuloksena on arvaamatonta käytöstä.

For example, email spam filtering uses statistical machine learning to determine good email vs spam email. By altering just 1 % of the training messages, the algorithm can be rendered ineffective. Attackers will use terms that appear in good emails and use them in spam email to force reclassification during retraining of fresh dataset.

- 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.

By introducing malicious data, attackers can cause the AI from operating correctly. In 2016 Microsoft unveiled Tay, a twitter bot that was intended to engage users between the ages of 18 and 24 through “casual and playful conversation”. Tay was “essentially a robot parrot with an internet connection” (Vincent, 2016). People started tweeting the bot with inappropriate remarks and within 24 hours of its release. This caused Tay to be turned into a foul-mouth racist, antifeminist, and Holocaust denier. According to Microsoft, Tay was built using relevant, modeled, clean and filtered data. However, when it went live and learnt from public data, its behavior drastically changed.

Lopputuloksena on, että malli ei toimi halutusti.

-Datan manipulointi

Data manipulation involves altering data within an ML model's training set to cause the model to misclassify data or behave in a predefined malicious manner in response to specific inputs. Techniques for manipulating training data include the following:

Adding incorrect data.

Removing correct data.

Injecting adversarial samples.

- Laukaisimilla luodut takaovet

In this scenario, attackers hide specific patterns or “triggers” within the model. When these patterns show up during real-world use, they make the model behave in unexpected ways. Imagine a sticker on a stop sign that confuses an autonomous vehicle, making it think it’s a yield sign and not a stop.

Gu et al. (2019), demonstrate that an attacker can create a backdoor vulnerability in the traffic sign classifier by inject multiple poisoned inputs during training. When the car approached a stop sign that had a special sticker, it would identify that stop sign as a speed limit sign.

In 2018, attackers were reporting toGmail, a massive amount of spam email as not spam, they were trying to containment training data in way that shifts the line between what the classifiers consider good data and what they consider to be bad data (Bursztein, 2018 & Constantin, 2021). By pushing corrupt data, attackers can manipulate malware detection systems to identify benign data as malicious and vis versa and alter spam filter systems used in mail to allow spam email as phishing, and corrupt a network instruction application where an unauthorized computer is now authorized.

Kyberturvallisuuden uhat

- Tekoälyä käytetään paljon muun muassa haittaohjelmien, roskapostin ja verkkoon tunkeutumisen havaitsemisessa.

Suojautuminen

* Koulutuksessa käytetyn tiedon lähteiden hallinta

Varmistaa lähteiden