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.