TEKOÄLY JA SEN HALLUSINOINTI

Ote Faktabaarin Tekoälyoppaasta: Generatiiviset tekoälymallit on koulutettu erittäin suuresta määrästä tekstiä, koodia, kuvia jne. Ne on usein poimittu Internetistä ja yleensä ilman omistajan lupaa. Itse asiassa emme tiedä tarkasti, mitä materiaalia koulutukseen on käytetty. Tekoälyn antamista vastauksista on vaikea päätellä, mihin aineistoon ne perustuvat lähdetietojen puuttumisen vuoksi. On tapauksia, joissa on todettu tekoälyn keksineen lähdemateriaaleja, joita ei ole oikeasti olemassa. Oppilaiden ja opiskelijoiden kanssa töitä tehdessä ja tekoälysovelluksia käyttäessä kannattaa muistaa kriittinen suhtautuminen AI:n antamiin lähteisiin ja osa tehtävää tulisikin olla lähteiden tarkistaminen. Se on toki aikaavievää, mutta opettaa lähdekritiikkiä ja tiedonhakua perusteellisella tavalla.

Markkinoille on tullut viimeisten kuukausien aikana lukuisia generatiivista tekoälyä hyödyntäviä ohjelmistoja ja sivustoja. Suurin osa niistä on kaupallisia ja kaikki niistä eivät ole välttämättä turvallisia ja luotettavia esimerkiksi tietosuojan kannalta. Monet organisaatiotovat rajoittaneet uusien ohjelmistojen käyttöä odottaen tarkempia tietoturvaselvityksiä niiden toiminnasta. Koulujen kannalta tilanne on hankala. Harvalla koulutuksenjärjestäjällä on tarvittavaa tietotaitoa tai resursseja tarkistaa uusien ohjelmistojen mahdollisia riskejä.

Generatiivisen tekoälyn (GenAI) käyttämät keinotekoiset neuroverkot ovat yleensä ”mustia laatikoita”, eli niiden sisäistä toimintaa ei voida tutkia. Tämän vuoksi ne eivät ole ”läpinäkyviä” eikä ole mahdollista selvittää, miten niiden tuotokset on syntyneet. Tekoälysovellukset voivat siis tuottaa odottamattomia tuloksia. Tämä läpinäkymättömyys on myös keskeinen syy GenAI-malleihin liittyviin luottamusongelmiin. Jos käyttäjät eivät ymmärrä, miten GenAI-järjestelmä on päätynyt tiettyyn tulokseen, he eivät todennäköisesti ole halukkaita ottamaan sitä käyttöön tai käyttämään sitä.

Siksi kaikkien tulisi tiedostaa, että GenAI-järjestelmät toimivat kuin mustat laatikot ja että sen vuoksi on vaikeaa ellei mahdotonta tietää, miksi tietty sisältö on luotu. Tuotokset voivat myös heijastaa tiettyjä kulttuurisia tai kaupallisia arvoja, jotka vääristävät tuotettua sisältöä. Nykyisellään vaikuttaa siltä[viii], että luotettavinta tietoa GPT-kielimallien opetusaineistoissa on Wikipedian sisältö, mutta sekään ei ole täysin neutraalia tai vapaata vinoumista, jotka periytyvät sen käyttäjäkunnalta.

Lähde: Faktabaarin Tekoälyopas