TEKOÄLYSTÄ TURVALLISTA

Antti Honkela

Apulaisprofessori, TkT

Tietojenkäsittelytieteen osasto

Helsingin yliopisto

Sisällys

Tekoälystä turvallista ... 3

Mistä on kyse ... 3

Tekoäly ja tietosuoja sekä yksityisyyden suoja ... 3

Tekoälyn tietoturva ... 4

Ihminen tekoälyn valvojana ja yhteistyökumppanina ... 5

Tekoäly muutosvoimana ... 6

Viestit päättäjille ja kansalaisille ... 7

Yksityiskohtaiset kommentit valtiopäiväasioista ... 8

U 28/2021 vp ... 8

Kainalolaatikko: Kansalaisten tietosuojan turvaava tietojen jako ... 8

Tekoälystä turvallista

Mistä on kyse

Yleisesti käytössä olevat tietokoneet ja muut tietotekniset laitteet ovat pohjimmiltaan koneita, jotka toistavat niille täsmällisesti ohjelmoituja tietojenkäsittelyoperaatioita erittäin nopeasti. Tekoälyjärjestelmät hyödyntävät näiden laitteiden nopeutta uudenkaltaisissa sovelluksissa, joissa järjestelmän toimintaa ei ole ohjelmoitu yksityiskohtaisesti suoraan, vaan esimerkiksi aiemmista esimerkeistä yleistäen tai monimutkaisen optimointiongelman ratkaisuna. Aiemmista esimerkeistä yleistämiseen perustuva koneoppiminen on mahdollistanut viimeisen 10 vuoden aikana valtavia edistysaskelia esimerkiksi konenäössä, puheentunnistuksessa ja luonnollisen kielen ymmärtämisessä ja tuottamisessa tietokoneella.

Keskityn kirjoituksessani kolmeen koneoppimiselle ominaiseen turvallisuuskysymykseen: opetusaineiston tietosuoja ja yksityisyyden suoja, tekoälyjärjestelmien tietoturva, sekä järjestelmien turvallisuus vuorovaikutuksessa ihmisten kanssa.

Koneoppimista käyttävien tekoälyjärjestelmien kehittämisessä tarvitaan suuria opetusaineistoja. Monet kiinnostavat tekoälyn sovellukset esimerkiksi terveydenhuollossa liittyvät ihmisiin, jolloin tarvittavat opetusaineistot sisältävät henkilötietoja. Näiden aineistojen keräämiseen ja käsittelyyn sekä niistä johdettujen järjestelmien käyttöön liittyy tietosuojariskejä ja ne saattavat loukata aineiston kohteiden yksityisyyttä.

Tekoälyjärjestelmät ovat tietojärjestelmiä. Kuten kaikissa tietojärjestelmissä, niissä voi olla tietoturvaheikkouksia. Tämä pitää huomioida, jos niitä käytetään turvattomassa ympäristössä, kuten Internetissä, tai epäluotettavilla tahoilla on mahdollisuus manipuloida niiden kehittämisessä käytettyjä opetusaineistoja tai muita syötteitä.

Tekoälyjärjestelmien ”älykkyys” poikkeaa merkittävästi ihmisen älykkyydestä. Ne voivat tehdä systemaattisesti ihmisen näkökulmasta järjettömältä vaikuttavia virheitä. Sovelluksissa, joissa virheistä voi olla merkittävää haittaa tai vaaraa, onkin tärkeää rakentaa mekanismeja näiden riskien ehkäisemiseksi esimerkiksi ihmisen valvontaa hyödyntäen.

Tekoäly ja tietosuoja sekä yksityisyyden suoja

Tyypillinen oppiva tekoälyjärjestelmä kuvaa syötteitä vasteille, esimerkiksi kuvan tiedoksi siinä näkyvistä liikennemerkeistä tai henkilön terveystiedot ennusteeksi nykyisestä tai tulevasta terveydentilasta.

Tällaisten järjestelmien rakentamiseen tarvitaan suuria esimerkkiaineistoja esimerkiksi suuren ihmisjoukon terveystiedoista ja havaitusta terveydentilasta, joista järjestelmä oppii näiden välisen yhteyden. Monissa kiinnostavissa sovelluksissa tarvittavat

aineistot sisältävät henkilötietoja kuten arkaluontoisia terveystietoja. Näiden käsittelyyn liittyy useita tietosuojariskejä, joita ei voida poistaa helposti esimerkiksi poistamalla tiedoista suorat tunnisteet.

1. Aineiston kerääminen yhteen tietokantaan lisää itsessään riskejä sen paljastumiseen, mikäli tietokannan suojaus pettää.
2. Aineiston osia voi paljastua järjestelmän opetuksen aikana opetukseen osallistuville ulkopuolisille tahoille.
3. Osia aineistosta voi olla mahdollista selvittää valmiiksi opetetusta järjestelmästä.

Kuvattuja riskejä voidaan ehkäistä erilaisin teknisin ratkaisuin. Järjestelmiä on mahdollista opettaa hajautettua dataa käyttäen, jolloin vältetään keskitettyyn tietokantaan liittyviä riskejä. Esimerkiksi Google ja Apple ovat opettaneet mobiilikäyttöjärjestelmiinsä käyttäjien kirjoituksia hyödyntäviä malleja, joiden opetuksessa käytetty data on pysynyt käyttäjien laitteilla eikä ole paljastunut edes mallin kehittäjille. Samaa periaatetta voidaan hyödyntää myös esimerkiksi yhdistettäessä terveystietoja eri maista, jos esimerkiksi lainsäädäntö ei salli datan luovutusta. **”Algoritmi tulee datan luo”** olisikin usein hyvä periaate parantamaan tietosuojaa.

Aineiston paljastumista opetuksen aikana voidaan estää salausmenetelmiä käyttämällä. Sopivilla menetelmillä on mahdollista opettaa järjestelmä salattua dataa käyttäen siten, että kehittäjä ei voi nähdä dataa. Hintana näillä ratkaisuilla on usein merkittävästi suurempi järjestelmän opettamiseen tarvittava laskenta-aika.

Aineiston paljastuminen valmiista järjestelmästä on hankalampi ongelma, johon perinteiset salausmenetelmät eivät pure. Esimerkiksi tulevaisuusvaliokunnassa esiintyneen ja tämän raportin myöhäisemmässä luvussa mainitun tekstiä tuottavan tekoälyn pohjalla olevan GPT-3 mallin edeltäjä GPT-2 on mahdollista houkutella toistamaan täydellinen kopio opetusaineistossa esiintyneistä henkilötiedoista. Sama ilmiö on havaittu muillakin testatuilla järjestelmillä. Ilmiötä voidaan estää hyödyntämällä differentiaalista tietosuojaa, mutta hintana on usein mallin tarkkuuden heikkeneminen. Ilmiön tarkempaan ymmärtämiseen tarvitaan lisätutkimusta. **Henkilötietoja tai muita luottamuksellisia tietoja käyttäen opetettujen tekoälyjärjestelmien jakamisessa on kuitenkin syytä olla varovainen.**

Tekoälyn tietoturva

Tekoälyjärjestelmät ovat tietojärjestelmiä ja alttiita samoille tietoturvariskeille kuin muutkin tietojärjestelmät. Erityisesti opetettaviin järjestelmiin liittyy lisäksi niille ominaisia riskejä, joista alla muutama esimerkki.

1. Opetusaineistoa manipuloimalla järjestelmän kehittäjä tai ulkopuolinen hyökkääjä voi upottaa järjestelmään takaoven, jota on käytännössä mahdoton havaita, mutta joka saa järjestelmän toimimaan valituissa tilanteissa hyökkääjän haluamalla tavalla. Esimerkiksi hakemusten käsittelyyn käytetty järjestelmä voitaisiin manipuloida hyväksymään tai hylkäämään aina hakemukset, joissa on tietty harmittomalta vaikuttava avainsana tai muu piirre.

2. Valmiista tekoälyjärjestelmästä on mahdollista löytää heikkouksia, joissa jopa ihmiselle huomaamattomat muutokset saavat järjestelmän toimimaan täysin eri tavalla. Tutkijat ovat kehittäneet esimerkiksi liikennemerkkiin kiinnitettäviä tarroja, jotka eivät erityisesti häiritse ihmistä, mutta joilla liikennemerkkien tunnistukseen käytetyt järjestelmät alkavat luulla stop-merkkiä nopeusrajoitukseksi. Vastaavia esimerkkejä on muiltakin aloilta, joissa on mahdollista tietoisesti kehittää esimerkkejä, jotka eivät hämäisi ihmistä, mutta joilla tekoälyä on mahdollista huijata.
3. Mikäli järjestelmä oppii adaptiivisesti toiminnan aikana, käyttäjät voivat ohjata sen toimintaa epätoivottuun suuntaan. Esimerkiksi Microsoftin käymistään keskusteluista oppinut Twitter-botti muuttui alle vuorokaudessa ystävällisestä aggressiiviseksi rasistiksi muiden käyttäjien manipuloinnin vuoksi.

Ylläkuvattuja riskejä voidaan ehkäistä yhdistelmällä teknisiä ja järjestelmätason ratkaisuja. **Tehokkain tapa opetusaineiston manipuloinnin ehkäisyyn on käytetyn aineiston tarkka laadunvalvonta. Valmiin järjestelmän käytössä vastaava ei ole yhtä helppoa, jolloin on tärkeää lisäksi vahvistaa järjestelmää niin, että se ei olisi niin altis huijausyrityksille** (vrt. stop-merkki-tarrat). Järjestelmien vahvistamiseen soveltuvat teknologiat ovat tällä hetkellä hyvin aktiivisen tutkimuksen kohteena, eikä ole vielä selvää mikä on lopulta mahdollista ja mikä ei. Kuten monessa turvallisuuskysymyksessä, tässäkin on riskinä hyökkääjien ja puolustajien välinen varustelukilpailu, jossa kehittyneempää puolustusjärjestelmää seuraa uusi vielä edistyneempi sen kiertävä hyökkäys.

Useimmat koneoppimiseen pohjautuvat tekoälyjärjestelmät eivät opi enää käytön aikana, mikä toisaalta helpottaa niiden toiminnan ymmärtämistä, toisaalta rajoittaa mahdollisuuksia sopeutua muuttuvaan ympäristöön. Käytön aikana tapahtuvan oppimisen turvallisuus on toistaiseksi hyvin nuori tutkimusala. Tehokkaita ratkaisuja sen ongelmiin saadaan todennäköisesti odottaa vielä useita vuosia.

Ihminen tekoälyn valvojana ja yhteistyökumppanina

Tekoälyjärjestelmiä käytetään enenevässä määrin sovelluksissa, joilla on merkittäviä vaikutuksia esimerkiksi ihmisten talouteen, hyvinvointiin ja terveyteen. Turvallinen tekoäly vaatii mekanismeja, joilla varmistetaan, että nämä vaikutukset eivät ole haitallisia. Nämä kysymykset ovat läheisessä yhteydessä tekoälyn vastuullisuuteen ja etiikkaan, mutta osa kysymyksistä liittyy myös turvallisuuteen: miten hallitaan autonomisten ajoneuvojen liikenneturvallisuusriskejä tai varmistetaan, ettei tekoälyn ehdottama hoitotoimenpide aiheuta tarpeetonta vaaraa?

Tekoälyjärjestelmille ja ihmisille on yhteistä se, että molemmat tekevät virheitä ja erehdyksiä. Näiden tekemät virheet ovat kuitenkin usein erilaisia, jolloin niitä voidaan ehkäistä koneen ja ihmisen yhteistyöllä. Matkalla on monta haastetta, kuten miten selitetään koneen toimintaa ihmiselle niin, että tämä voi sitä arvioida ja havaita mahdollisia ongelmia; miten kehittää järjestelmiä, jotka ovat helpommin ihmisen ymmärrettäviä menettämättä niiden tehokkuutta; miten rakennetaan ihmiselle mahdollisuus valvoa koneen toimintaa niin, että tehtävä pysyy mielekkäänä ja ihminen pystyy oikeasti

puuttumaan tarpeen vaatiessa riittävän nopeasti; miten järjestetään ihmisohjaus menettämättä liikaa koneen tehokkuushyötyjä; miten koulutetaan ihmisiä seuraamaan koneiden toimintaa tehokkaasti?

Kuten tavalliset tietojärjestelmät, tekoälyjärjestelmät eivät opi virheistään, ellei niitä ole erityisesti siihen suunniteltu. Yleisin tapa korjata virheitä on uuden version julkaisu ja käyttöönotto, mikä rajoittaa mahdollisuuksia nopeaan reagointiin, erityisesti jos käyttäjä ja kehittäjä ovat erillisiä tahoja. Aiemmin mainittu aktiivinen oppiminen käytön aikana mahdollistaisi teoriassa nopeamman reagoinnin, mutta näiden menetelmien turvallisuudessa on vielä paljon avoimia ongelmia.

Tekoäly muutosvoimana

Erilaiset tekoälyjärjestelmät ovat viime vuosina tulleet enenevässä määrin osaksi ihmisten arkea erilaisissa nettipalveluissa ja elektronisissa laitteissa. Niiden yleistymisen voi olettaa jatkuvan ja laajenevan uusille aloille. Paljon potentiaalia on esimerkiksi terveydenhuollossa. Toimiessaan nämä järjestelmät sujuvoittavat arkea ja tehostavat palveluita, mutta samalla ne luovat myös uusia turvallisuusuhkia.

Tekoälyjärjestelmät ovat ensisijaisesti tietojärjestelmiä ja niihin pätevät monet samat lainalaisuudet. Perinteisten tietojärjestelmien turvallisuudesta muodostuu valitettavan usein ”kisa pohjalle”, jossa järjestelmien rakentajat kilpailevat säästämällä turvallisuudesta leikatakseen kustannuksia, koska ostajat eivät osaa sitä vaatia eivätkä ole valmiita siitä maksamaan. **Yksi osa ongelmaa on se, että turvallisuuden puutteet koituvat usein muiden vahingoksi. Kuinka moni auton ostaja olisi vapaaehtoisesti valmis maksamaan enemmän autosta, joka olisi kolaritilanteessa turvallisempi muille, esimerkiksi jalankulkijoille?**

Turvallisuuteen liittyvissä sovelluksissa hyökkääjän ja puolustajan välille muodostuu helposti varustelukierre, jossa paremman puolustuksen hyöty häviää paremman hyökkäyksen edessä, mikä edellyttää vielä parempaa puolustusta ja niin edelleen. Tällaisia kierteitä voi syntyä myös tekoälyn turvaa murrettaessa ja puolustettaessa. Tekoäly on vahva työkalu kisan molemmilla puolilla, jolloin molempien osapuolten voi olla käytännössä pakko hyödyntää sitä pärjätäkseen kilpailussa. Tieteellinen perustutkimus voi joissain tilanteissa tarjota ulospääsyn varustelukierteestä. **Esimerkiksi differentiaalista tietosuojaa käyttäen voidaan todistaa matemaattisesti, että mikään hyökkääjä ei voi loukata käyttäjien tietosuojaa, ei vaikka tämä käyttäisi mitä tahansa tulevaisuuden teknologiaa.** Vastaavasti tekoälyjärjestelmien huijaamista tai kiertämistä vastaan (vrt. stop-merkki-tarrat) kehitetään teknologioita, joilla voidaan todistaa mallin kestävän varmasti tiettyjä manipulaatioita.

Turvallisen tekoälyn teknologioiden kehitys tuottaa todennäköisesti tulevaisuudessa helppokäyttöisiä työkaluja, joilla järjestelmien turvallisuutta voidaan vahvistaa ja varmistaa. **Turvallisen tekoälyn käytössä tullaan kuitenkin todennäköisesti törmäämään vaikeisiin arvovalintoihin, koska valinnoilla on hintansa.** Tietosuojaa kunnioittava tai manipulaatioita kestävä järjestelmä tuottaa usein optimiolosuhteissa vä-

hemmän tarkkoja ennusteita, jolloin järjestelmän tuottama hyöty pienenee. **Mikä pahempaa, tämä tarkkuuden lasku vaikuttaisi kohdistuvan helposti voimakkaampana vähemmistöihin ja yleensä enemmistöstä poikkeaviin yksilöihin, mikä voi loukata yhdenvertaisuutta.** Näiden eri arvojen yhteensovittamisesta tarvitaan vahvaa yhteiskunnallista keskustelua, ettei niiden ratkaisu jää yksin teknologian kehittäjien vastuulle. Yhteensovittamisessa on paljonkin pelivaraa, koska turvallisuus ei ole päällä/pois -tyyppinen ominaisuus vaan useimmissa teknologioissa suojan tasoa ja siten kustannuksia ja sivuvaikutuksien voimakkuutta on mahdollista säätää.

Viestit päättäjille ja kansalaisille

Tietoturvan ja -suojan sääntely on tietoyhteiskunnan ympäristönsuojelua. Koska kustannukset kohdistuvat ulkopuolisiin, markkinat eivät sitä ratkaise, vaan tarvitaan ulkopuolista ohjausta ja sääntelyä.

Turvallisen ja eettisen tekoälyn toivottavat ominaisuudet voivat olla ristiriitaisia, jolloin yhden toteuttaminen heikentää muita. Ominaisuudet eivät ole mustavalkoisesti päällä/pois, vaan niiden välistä tasapainoa voi yleensä säätää tarkemmin. Tarvitaan sopivasti vapautta ja sääntöjä löytää kuhunkin sovellukseen paras ratkaisu. Samalla pitää varmistaa, että tasapainoa muodostettaessa huomioidaan myös muiden kuin järjestelmän kehittäjän ja omistajan etu.

Valmiista tekoälyjärjestelmästä on mahdoton verifioida, toimiiko se joka tilanteessa oikein eikä esimerkiksi sisällä piilotettua takaovea. Luotettavuuden varmistamiseksi on tärkeää lisäksi verifioida prosessi, jolla järjestelmä on rakennettu.

Tekoälyjärjestelmät eivät ole erehtymättömiä, mutta eivät ole ihmisetkään. Näiden kahden tekemät virheet ovat luonteeltaan hyvin erilaisia. Tärkeissä sovelluksissa käytettäviä tekoälyjärjestelmiä tulee valvoa tavalla, joka on tarkoituksenmukainen kyseiseen sovellukseen, eikä turhaan nosta valvonnan kustannuksia mahdollisten virheiden haittoja suuremmiksi.

Datan luovuttamisen sijaan kannattaa suosia järjestelmiä, joissa koodi tulee valvotusti datan luo. Esimerkiksi terveystietojen ulkopuolisille luovuttamisen sijaan olisi parempi suosia ratkaisuja, joissa dataa ei luovuteta vaan sen tarvitsijat pääsevät käsittelemään tarvitsemiaan osia turvallisesti suojatussa ympäristössä.

Datan keskittämisen sijaan kannattaa mahdollisuuksien mukaan edistää datan hajauttamista tietosuojariskien laajuuden pienentämiseksi. Suunniteltu keskitetty genomikeskus on esimerkki erittäin suuresta riskistä, jonka hyödyistä on erimielisyyttä.

Tekoälyn tietoturvan ja tietosuojan osaamista tulee ylläpitää ja kehittää aktiivisesti kaikilla tasoilla. Turvaa ei usein voi lisätä jälkikäteen valmiiseen järjestelmään, joten kehittäjien pitää osata rakentaa se mukaan alusta asti, mikä vaatii laajaa koulutusta. Toisaalta tarvitaan uusien turvallisempien menetelmien tutkimusta. Vain panostamalla alan perustutkimukseen Suomessa voidaan varmistaa paras osaaminen ja korkein opetus uusista menetelmistä.

Yksityiskohtaiset kommentit valtiopäiväasioista

U 28/2021 vp

U-kirjelmän kohdassa 7.2 kiinnitetään huomiota ehdotetun tekoälyasetuksen artiklan 10(5) sallimaan arkaluonteisten henkilötietojen käsittelyyn tekoälyjärjestelmissä esiintyvien vinoumien korjaamiseksi.

Kyseinen kohta on yksi tekoälyasetuksen parhaista puolista, ja **vastaavan säädöksen sovellusalaa kannattaisi pikemmin harkita laajennettavaksi korkean riskin tekoälyjärjestelmistä kaikkiin ihmisiä koskevia päätöksiä tekeviin järjestelmiin**.

Artiklan 10(5) taustalla on tieteellinen tosiseikka, että tekoälyjärjestelmän vinoumia on mahdotonta estää, ellei järjestelmän kehittäjä pysty käsittelemään tarvittavia arkaluontoisia henkilötietoja. Tutkimuskirjallisuudessa on lukuisia esimerkkejä, joissa tekoälyjärjestelmä, joka opetetaan ilman näitä tietoja, tuottaa vinoutuneita päätöksiä. Näiden vinoumien korjaaminen on mahdollista ainoastaan tarvittavia arkaluonteisia henkilötietoja käyttämällä.

Kainalolaatikko: Kansalaisten tietosuojan turvaava tietojen jako

Yhdysvalloissa järjestetään 10 vuoden välein väestönlaskenta (US Census). Laskennan tuloksena julkaistaan tilastoja eri ikä-, etnisiin ym. ryhmiin kuuluvista henkilöistä tarkalla maantieteellisellä jaolla. Näitä tilastoja käytetään erilaisiin hallinnollisiin ja tutkimustarkoituksiin. Tietoja on aiemmin suojattu mm. korvaamalla pieniä ihmismääriä koskevia tietoja synteettisillä arvoilla. Väestölaskentaviraston tutkimuksen mukaan näistä suojatoimista huolimatta huomattava osa väestöstä voitiin tunnistaa julkaistuista tiedoista yhdistämällä niitä muihin julkisesti saatavilla oleviin tietoihin.

Uusin vuoden 2020 US Census on ottanut käyttöön tietojenkäsittelytieteen yksityisyyden turvaavien teknologioiden tutkimukseen perustuvan differentiaalisen tietosuojan tietojen suojaamisessa. Differentiaalisen tietosuojan käyttö vähentää huomattavasti tunnistamisriskiä ja parantaa siten selvästi kansalaisten yksityisyyden suojaa. Sen käyttöä on silti myös kritisoitu, koska tietosuojan turvaaminen vaatii lisäämään tuloksiin enemmän satunnaisuutta, mihin tietojen käyttäjät eivät ole tottuneet ja mikä saattaa haitata joitain totuttuja tietojen käyttötarkoituksia.

Differentiaalisen tietosuojan kaltaisten menetelmien käyttö avaa laajemminkin mahdollisuuksia yksityisyyden turvaavaan tietojen jakamiseen, joten Yhdysvalloissa saatuja kokemuksia on syytä seurata. Yksi erityisen kiinnostava mahdollisuus on differentiaalisen tietosuojan yhdistäminen ns. synteettiseen dataan, jossa esimerkiksi koneoppivan tekoälyn opettamiseen voidaan luoda todellista vastaava keinotekoinen aineisto, joka ei todistettavasti voi loukata kenenkään yksityisyyttä. Näillä menetelmillä on kuitenkin rajoitteensa, ja lisätutkimusta tarvitaan niiden potentiaalin ymmärtämiseksi.