Algoritmien valta sekä neutraalius ja puolueellisuus koneellisessa päätöksenteossa

Lausunnon antaja:
KTT Joni Salminen, Turun yliopiston kauppakorkeakoulu

Sisältö: Tämä dokumentti avaa algoritmien valtaan ja koneelliseen päätöksentekoon liittyviä keskeisiä yhteiskunnallisia ulottuvuuksia. Se sisältää seitsemän kohtaa.

A. Neutraali algoritmi: Mikäli algoritmin toimintaperiaate on läpinäkyvä ja yleisesti hyväksytty, voidaan sen katsoa olevan prosessin puolesta neutraali. Algoritmi voi kuitenkin silti tuottaa puolueellisia tuloksia, erityisesti koneoppimisen kautta. Monet ongelmat liittyvät koneiden tapaan oppia, joka tapahtuu todennäköisyyslaskennan kautta. Koneellisessa päätöksenteossa voidaan erottaa ainakin kolme puolueellisuuden lähdettä: (1) algoritmit ovat ihmisten suunnittelemia, joten algoritmin kehittäjän puolueellisuus saattaa tietoisesti tai tiedostamatta siirtyä niiden toimintaperiaatteisiin. Toiseksi etenkin informaatioalgoritmit ovat alttiita (2) käyttäjien hyväksikäytölle, kuten järjestäytyneelle manipuloinnille. Kolmanneksi (3) datalähteet saattavat olla vääristyneitä tai sisältää puolueellisuutta, jolloin myös koneen tulkinta vääristyy ja tehdyt päätökset voivat olla epäreiluja.

B. Reiluuden kriteerit:
Eettisesti hyvän algoritmin ominaisuudet kulminoituvat käsitteeseen FATML, joka tulee sanoista Fair (reilu), Accountable (vastuullinen), Transparent (läpinäkyvä) ja Machine Learning (koneoppiminen), eli kyseessä on Reilu, Vastuullinen ja Läpinäkyvä Koneoppiminen (RVLK näin suomeksi).

C. Esteitä läpinäkyvyydelle: ACM:n mukaan[1] algoritmien läpinäkymättömyyteen on olemassa kolme pääsyytä: (1) tekniset syyt (ts. algoritmin toimintaa ei ole helppo selittää), (2) taloudelliset syyt (ts. liikesalaisuudet estävät algoritmien toiminnan paljastamisen), ja (3) sosiaaliset syyt (ts. tietojen paljastaminen voi rikkoa odotuksia yksityisyydensuojasta). Lisään tähän neljännen: (4) tietojen paljastaminen voi rikkoa lakeja yksityisyydensuojasta tai palvelun itse määrittelemiä käyttöehtoja.

D. Algoritmien läpinäkyvyyden ja vastuullisuuden ohjenuorat (ACM:n mukaan):

- **Tietoisuus:** Analyyttisten järjestelmien omistajien, suunnittelijoiden, rakentajien, käyttäjien ja muiden sidosryhmien tulisi olla tietoisia mahdollisista ennakkoluuloista, jotka liittyvät niiden suunnitteluun, käyttöönottoon ja käyttöön sekä potentiaalisiin haittoihin, joita niiden puolueellisuus voi aiheuttaa yksilöille ja yhteiskunnalle.
- **Pääsy:** Lainsäätäjiä olisi kannustettava sellaisten mekanismien käyttöönottoon, jotka mahdollistavat koneellisten päätösten kyseenalaistamisen ja oikeussuojan antamisen yksilöille ja ryhmille, joita koneelliset päätökset mahdollisesti haittaavat.

[1] https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

- **Vastuullisuus:** Organisaatioiden tulisi olla vastuussa käytettävien algoritmien tekemistä päätöksistä, vaikka ei olisikaan mahdollista selittää yksityiskohtaisesti, miten algoritmit tuottavat tuloksensa.
- **Selitys:** Algoritmista päätöksentekoa käyttäviä järjestelmiä ja organisaatioita kannustetaan tuottamaan selityksiä sekä algoritmin suorittamista toimenpiteistä että yksittäisistä päätöksistä. Tämä on erityisen tärkeää julkishallinnon yhteydessä.
- **Tietojen alkuperä:** Algoritmien kehittäjien tulisi dokumentoida, kuinka koneoppimismallien harjoitteluaineisto kerättiin, sekä arvioida ihmisen tai koneellisen tiedonkeruuprosessin aiheuttamia mahdollisia vääristymiä. Tietojen julkinen tarkastelu tarjoaa mahdollisuuden korjata mallien aineistosta johtuvia virheitä. Kuitenkin huoli yksityisyyden suojasta, liikesalaisuuksien suojelemisesta tai sellaisten toiminnallisuuksien paljastamisesta, jotka voivat antaa pahaa tarkoittaville toimijoille mahdollisuuden hyväksikäyttää järjestelmää, voivat oikeuttaa rajoittamaan pääsyn päteviin ja luotettaviin henkilöihin.
- **Tarkastettavuus:** Mallit, algoritmit, tiedot ja päätökset on dokumentoitava siten, että ne voidaan ulkopuolisen tahon toimesta tarkastaa tapauksissa, joissa haittaa epäillään.
- **Validointi ja testaus:** Organisaatioiden tulee käyttää uskottavia menetelmiä malliensa vahvistamiseen ja dokumentoida nämä menetelmät ja tulokset. Erityisesti niiden tulisi tasaisin väliajoin suorittaa testejä arvioidakseen ja määrittääkseen, aiheuttaako malli yhteiskunnallisia haittoja. Koneellista päätöksentekoa soveltavia organisaatioita kannustetaan tekemään tällaisten testien tulokset julkiseksi.

E. Euroopan tilanne: Tarkastellaan koneellista päätöksentekoa eurooppalaisesta näkökulmasta. Käytännössä kaikki ns. superalustat (alusta, jolla on yli sata miljoonaa globaalia käyttäjää) ovat yhdysvaltalaisomistuksessa. Nämä kaupalliset toimijat, joiden algoritmien tarkka toiminta on tuntematonta niin käyttäjille kuin lainsäätäjille, päättävät minkälaista tietoa eurooppalaisille kansalaisille päivittäin näytetään. Eurooppalaiset viettävät esimerkiksi yhteensä yli 60 miljoonaa tuntia päivässä Facebookissa ja tekevät yli 30 miljardia hakua Googlessa kuukausittain. Nuorempien sukupolvien mediakulutus on vielä enemmän verkkopainotteista, ja useille verkkomediat muodostavat tärkeimmän informaatiolähteen.

On siis poikkeuksellisen selvää, että amerikkalaisilla yrityksillä on superalustojen kautta suuri valta yleisen mielipiteen muodostumisessa Euroopassa. Koska vallan luonteeseen kuuluu väärinkäytön mahdollisuus, eurooppalaisten on kriittistä vähintäänkin ymmärtää, mistä algoritmien toiminnassa on kyse. Tämä ymmärrys voidaan kytkeä mm. lainsäädäntöön, jotta superalustojen toimintaa voidaan valvoa ja tarvittaessa kontrolloida.

F. Sääntelyn merkitys: Sääntelyn lisääminen kotimarkkinoilla tai EU-tasolla ei kuitenkaan ole ainoa tapa vastata. Sääntelyn sijaan tarvitaan kilpailukykyä – eurooppalaisten pitää kyetä haastamaan amerikkalaiset yritykset tekoälyn kehittämisessä. Samaa tekee myös Kiina, joka esimerkiksi tukee Baidu-hakukoneen tekoälykehitystyötä – näiden keinojen tavoitteena on pysyä mukana tekoälykilvassa. Tulevaisuuden maailmassa voi olla useita eri tekoälyratkaisuja, jotka kilpailevat toisiaan vastaan. Vielä toistaiseksi hyvää on ollut se, että isot toimijat jakavat

työkalujaan avoimen lähdekoodin alla (esim. Googlen Tensorflow ja Facebookin Torch), mutta dataa ne varjelevat tarkoin ja nämä työkalut ovat hyödyttömiä ilman dataa.

G. Toimijoiden välisen yhteistyön lisääminen: Vahvat auditoinnin standardit ja tekoälyn toiminnan ymmärtäminen on tarpeellista. Kuitenkaan tällä hetkellä ei ole olemassa vakiintuneita käytänteitä tekoälyn vaikuttavuuden arvioimiseen. Tutkimushankkeet, jotka keskittyvät tekoälyn yhteiskunnallisiin vaikutuksiin ovat varsin hajanaisia niin Suomessa kuin Euroopassa. On niin julkisrahoitteisia kuin yksityisrahoitteisia hankkeita – tekoälyn valtaa tutkitaan tällä hetkellä eniten Yhdysvalloissa, mutta myös Euroopassa (ja Suomessa) on useita hankkeita käynnissä. Näiden välistä koordinaatiota tulisi kuitenkin lisätä, jotta standardien luonti olisi mahdollista.