Datan ja tekoälyn eettiset periaatteet

Mikko Niemi 2022

Helsinki

Johdanto

Kaupunki haluaa olla vastuullinen edelläkävijä tekoälyn hyödyntämisessä. Tekoäly perustuu dataan, joka voi olla esimerkiksi mittaustuloksia tai tekstiä. Data kuvaa tietoa esimerkiksi henkilöistä, sijainnista tai palveluiden tilasta. Dataa muodostuu suuressa osassa kaupungin toimintaa, esimerkiksi liikenteen järjestelmissä tai kaupungin palveluita käytettäessä. Tekoäly on työkalu, jolla kaupungin datan hyödyntämistä voi mahdollistaa, tukea ja tehostaa.

Tekoälyllä tarkoitetaan tietokoneohjelmia, jotka kykenevät toimintoihin, joiden usein ajatellaan vaativan ihmisen älykkyyttä. Näihin lukeutuu esimerkiksi tulevaisuuden ennustaminen menneisyyden havaintojen pohjalta, ongelmanratkaisu kohti jotain tiettyä päämäärää, päätöksenteko annettujen parametrien pohjalta, hahmojen tunnistaminen kuvista, tiivistelmän luominen tekstistä ja kielen kääntäminen. Tekoälyratkaisuja on ollut olemassa jo pitkään mutta ala kehittyy nyt voimakkaasti ja erilaisia sovelluksia otetaan aktiivisesti käyttöön.

Datan hyödyntäminen ja tekoälyjärjestelmät mahdollistavat aiempaa tehokkaamman automatisoinnin ja palveluiden laadun parantamisen, ja siksi niitä halutaan hyödyntää enemmän. Tähän liittyy kuitenkin eettisiä kysymyksiä. Helsingin kaupunki haluaa itse käyttää dataa ja tekoälyä vastuullisesti ja eettisesti, ja toimia suunnannäyttäjänä muille. Tästä tavoitteesta syntyivät tekoälyn ja datan eettiset periaatteet. Näitä kahdeksaa periaatteetta noudattamalla kaupunki minimoi dataan ja tekoälyyn liittyviä riskejä.
Datan ja tekoälyn eettiset periaatteet sisältävät itse eettiset periaatteet sekä kysymyslistauksen jokaisen periaatteen toteutumisen käytännön arvioimiseksi.

Tausta

- Järjestimme kansainvälisen tekoälyn etiikkaa käsittelevän seminaarin 2019
- Toteutimme tekoälyrekisterin yhteistyössä Amsterdamin kanssa 2020
- Olimme mukana toteuttamassa HY:n Ethics of AI-kurssia
- Laadimme sopimusliitteen tekoälyhankintoihin huomioiden eettiset periaatteet, pohjana Amsterdamin vastaava dokumentti

Prosessi

- Periaatteet perustuvat meta-analyyseihin tekoälyn eettisistä periaatteista
- Moniammatillinen asiantuntijaryhmä eri toimialoilta muotoili periaatteet
- Digitaalisen johtoryhmän lähetekeskustelu
- Sidosryhmien osallistaminen käynnissä
- Digitaalisen johtoryhmän linjaus periaatteista
- Viimeinen askel kansliapäällikön päätös

Sidosryhmät

- Ensisijainen kohderyhmä on koko ammatillinen kenttä: projektipäälliköt, datatieteilijät, ohjelmoijat ja muut tekoälyn parissa työskentelevät
- Muu kaupungin henkilöstö
- Kaupunkilaiset
- Poliitikot
- Media
- Akatemia

Osallistaminen

- Kaupungin koneoppimisverkosto
- Akatemia (FCAI)
- HRI Loves Developers
- Digitaalinen johtoryhmä
- Henkilöstö
- Kaupunkilaiset
- Tekoälytoimittajat
- Tasa-arvo- ja yhdenvertaisuustoimikunta

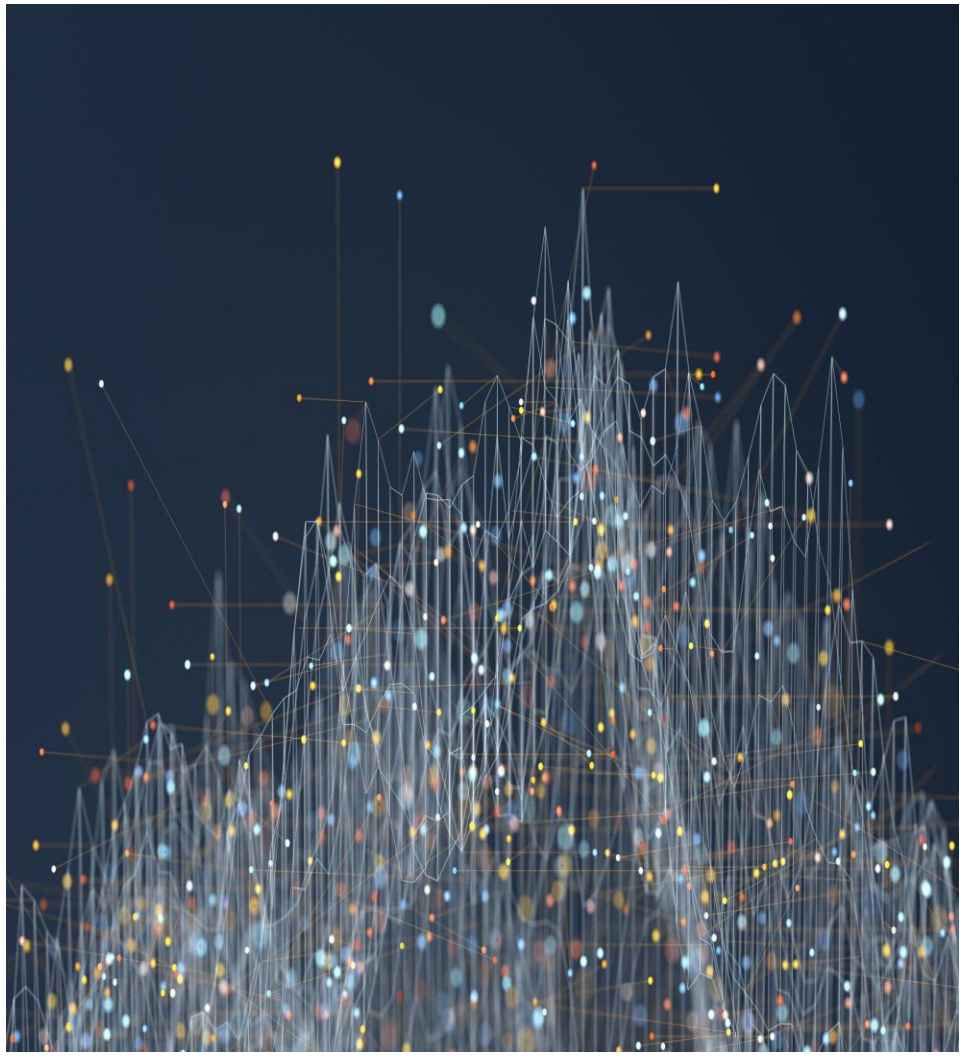
Jalkauttaminen

- Toimialojen osallistaminen valmistelutyöhön
- Vuorovaikutustilaisuudet
- Digitaalisen johtoryhmän käsittelyt
- Päätöksen jälkeen erillinen viestintä
- Periaatteiden kommunikointi verkostojen kautta

Periaatteet

Datan ja tekoälyn eettiset periaatteet

- Ihmislähtöisyys
- Läpinäkyvyys
- Selitettävyys
- Oikeudenmukaisuus ja yhdenvertaisuus
- Vastuu ja luottamuksen ylläpitäminen
- Yksityisyys
- Turvallisuus
- Ihmisen kontrollissa



Periaatteet yksi kerrallaan

Ihmislähtöisyys

- **Mitä?** Kehitämme dataa ja tekoälyä hyödyntäviä palveluita ihmisten hyväksi ja heitä kuunnellen.
- **Miksi tärkeä?** Julkisena toimijana palvelemme ihmisiä ja yhteisöjä.

Läpinäkyvyys

- **Mitä?** Viestimme mahdollisimman ymmärrettävästi, miten ja missä hyödynnämme dataa ja tekoälyä. Pyrimme aktiivisesti jakamaan kaiken keskeisen tiedon, ellei lainsäädäntö sitä estä. Tuomme datan tunnistetut puutteet esille. Mikäli mahdollista, jaamme myös lähdekoodin.
- **Miksi tärkeä?** Ihmisillä tulee olla mahdollisuus saada tietoa datan ja tekoälyn hyödyntämisestä ymmärrettävässä muodossa, jotta he voivat arvioida kaupungin toimintaa. Luottamus voi lisääntyä vain avoimuuden kautta.

Selitettävyys

- **Mitä?** Jos datan ja tekoälyn hyödyntämiseen liittyy merkittäviä riskejä, pystymme selittämään yksittäisen tuloksen tai algoritmin yleisen toimintalogiikan ymmärrettävästi.
- **Miksi tärkeä?** Selitettävyys edistää järjestelmän vaikutusten hallintaa. Jos tekoälyn toiminnalla on oikeusvaikutuksia, selitettävyys auttaa varmistamaan asiakkaan oikeuksien toteutumisen.

Oikeudenmukaisuus ja yhdenvertaisuus

- **Mitä?** Datan käytön ja tekoälyratkaisujen lähtökohtana on jokaisen ihmisarvon ja oikeuksien kunnioittaminen. Huomioimme mahdolliset vinoumat sekä datassa että algoritmeissa, ja puutumme niihin tarvittaessa.
- **Miksi tärkeä?** Tekoälyjärjestelmät voivat toistaa ja voimistaa olemassa olevia epätasa-arvoisia rakenteita.

Vastuu ja luottamuksen ylläpitäminen

- **Mitä?** Osoitamme jokaiselle tekoälyä hyödyntävälle palvelulle vastuutahon, johon asiakkaillamme on mahdollisuus saada yhteys. Vastuutaho huolehtii, että datan ja tekoälyn eettisiä periaatteita noudatetaan.
- **Miksi tärkeä?** Haluamme, että asiakas voi luottaa kaupungin toimintaan dataa ja tekoälyä hyödynnettäessä.

Yksityisyys

- **Mitä?** Käsittelemme henkilötietoja huolellisesti ja tietoturvallisesti järjestelmän koko elinkaaren ajan.
- **Miksi tärkeä?** Yksityisyydestä huolehtiminen mahdollistaa datan hyödyntämisen turvallisesti ja luotettavasti.

Turvallisuus

- **Mitä?** Dataa ja tekoälyä hyödyntävät järjestelmät ovat hyvin suojattuja ja hallittuja. Tunnistamme ja minimoimme mahdolliset riskit. Riskienhallinta saattaa rajoittaa mahdollisuutta jakaa tietoa järjestelmistä.
- **Miksi tärkeä?** Minimoimme mahdolliset turvallisuusriskit, jotta kaupunkilainen voi hyötyä datan ja tekoälyn käytöstä.

Ihmisen kontrollissa

- **Mitä?** Joissain tilanteissa järjestelmän riskien hallinta vaatii korkeampaa ihmisen kontrollin tasoa. Tällöin varmistamme, että vastuussa oleva henkilö pystyy seuraamaan ja valvomaan järjestelmän toimintaa sekä tarvittaessa puuttumaan siihen.
- **Miksi tärkeä?** Tekoälyjärjestelmä ei voi oppia kaikkia mahdollisia tilanteita. Virhepäätelmän tai teknisen vian riski on aina olemassa. Ihminen kontrollissa mahdollistaa yllättäviin tilanteisiin ja vaikutuksiin reagoimisen.

Keskustelu

Keskustelu