DIGI- JA VÄESTÖTIETO-VIRASTO

Turvallisen kehittämisen opas

Tekoälyjärjestelmien kehittäminen

9.6.2023

Puh. **0295 536 000** (vaihde) | Yhteystiedot, katso **www.dvv.fi**

Sisällysluettelo

1 Turvallinen tekoälyjärjestelmien kehittäminen

1.1 Johdanto

Tekoälyllä tarkoitetaan tietojärjestelmiä tai laitteita, jotka osaavat toteuttaa älykkäänä pidettäviä toimintoja. Tällaisia älykkäitä toimintoja ovat esimerkiksi puheentunnistus, kuvantunnistus, erilaiset suositteluratkaisut tai automaattisen päätöksenteon järjestelmät. Yhä useammin tekoälyjärjestelmien toteuttamat älykkäät toiminnot toteutetaan koneoppimisen avulla: älykäs toiminta opitaan datasta.

Turvallinen kehittäminen edellyttää, että hyödynnettävän datan käsittely on turvallista ja lainmukaista. Tekoälyjärjestelmien toteutuksessa on varmistettava kansalaisten perusoikeuksien toteutuminen. Tekoälyä hyödyntävät tekniset laitteet eivät saa aiheuttaa fyysistä tai henkistä uhkaa terveydelle tai hyvinvoinnille. Datan ja tekoälyn käytöllä voi olla myös hitaasti kehittyviä pitkän aikavälin vaikutuksia, joita on tarpeellista pohtia ennakoivasti (1).

Tämä opas on tarkoitettu kaikille tekoälystä kiinnostuneille, mutta erityisesti tekoälyhankkeiden vastuuhenkilöille, toteuttajille ja hankinnoista vastaaville. Oppaan tarkoitus on tuoda kattavasti esiin tekoälyjärjestelmien toteutukseen liittyviä haasteita, jotka voivat vaarantaa kansalaisten tai käyttäjien turvallisuuden tai perusoikeudet. Toiveena on, että tekoälyhankkeita suunniteltaessa ja toteutettaessa huomioidaan etukäteen kaikki tarpeelliset turvallisuutta edellyttävät näkökulmat.

1.2 Vaatimukset turvalliselle tekoälyjärjestelmälle

Turvallisen tekoälyjärjestelmän toteuttaminen edellyttää useiden eri näkökulmien huomioimista suunnittelussa ja toteutuksessa. Alla olevassa taulukossa on esitetty tärkeimmät turvallisen järjestelmän vaatimukset sekä ohjeita turvallisuuden lisäämiseksi.

Taulukko 1: Vaatimukset turvalliselle tekoälyjärjestelmälle.

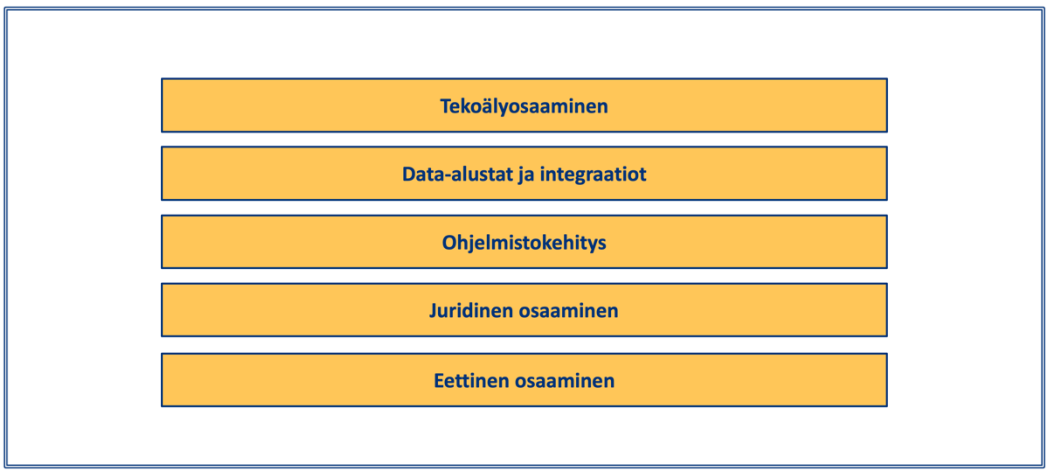
Vaatimus	Ohjeita
Järjestelmä huomioi tietoturvan ja tietosuojan	Varmista, että luottamus, eheys, saatavuus -periaatteet toteutuvat. Dataan on pääsy vain oikeilla henkilöillä. Tiedon keruussa ja tallentamisessa noudatetaan lainsäädäntöä, erityisesti tietosuoja-asetusta.
Järjestelmä toteuttaa kansalaisten perusoikeudet	Älykkäät järjestelmät ja laitteet kohtelevat kaikkia kansalaisia lainmukaisesti, kenenkään oikeuksia vaarantamatta. Hankkeen valmistelussa on perehdyttävä datan ja tekoälyn hyödyntämisen kannalta relevanttiin lainsäädäntöön.
Järjestelmä on turvallinen käyttäjille ja sen vaikutuspiirissä oleville	Älykkäät järjestelmät ja laitteet eivät aiheuta ihmisille vaaraa tai riskejä. Tämä koskee erityisesti kriittistä infrastruktuuria esim. liikennettä. Hankkeessa suunnitellaan ja toteutetaan toiminnan laadunvarmistuksen prosessit.

Ratkaisu on pitkällä aikavälillä eettisesti kestävä	Hankkeessa varmistetaan eettisten periaatteiden huomioiminen kaikissa vaiheissa. Älykkäiden järjestelmien toteutuksessa huomioidaan myös pitkän aikavälin vaikutukset yhteiskuntaan, ja tavoitellaan positiivisia hyötyjä.

Lisätietoja tietoturvasta ja tietosuojasta löytyy turvallisen sovelluskehityksen käsikirjasta (2), Virta-arkkitehtuurista (3), sekä EU:n yleisestä tietosuoja-asetuksesta (4).

Lisätietoa tekoälyyn liittyvästä lainsäädännöstä löytyy EU:n ehdotuksesta tekoälyasetukseksi (5). Lisätietoa eettisen tekoälyn toteuttamiseen löytyy EU:n luotettavan tekoälyn eettisistä ohjeista (6).

1.3 Tekoälyhankkeissa tarvittavat valmiudet



Kuva 1: Tekoälyhankkeita tukevat valmiudet.

Dataa ja tekoälyä hyödyntävä turvallinen kehittäminen vaatii organisaatiolta erilaisia valmiuksia, jotka tukevat hanketta sen eri vaiheissa. Näiden valmiuksien kehittäminen hankkeen aikana ei välttämättä onnistu. Onnistuneiden hankkeiden taustalla on jatkuva ja systemaattinen perusvalmiuksien kehittäminen.

Tekoälyjärjestelmien merkittävä riippuvuus datasta sekä kyvykkyys päätöksenteon automatisointiin luovat entistä voimakkaamman riippuvuuden teknologisen osaamisen, juridisen osaamisen ja eettisen osaamisen välille. Tarvittavat valmiudet on lyhyesti kuvattu alla olevassa taulukossa.

Taulukko 2: Organisaatioiden valmiudet tekoälyhankkeiden tukemiseksi.

Valmius	Kuvaus
Tekoälyosaaminen	Dataa ja tekoälyä hyödyntävien organisaatioiden tulee huolehtia henkilöstön kouluttamisesta. On suositeltavaa, että datan ja tekoälyn toimintaperiaatteet ymmärretään organisaatiossa laajasti, myös ei-teknisissä asiantuntijatehtävissä sekä ylimmässä johdossa. On suotavaa muodostaa yhteinen ymmärrys ja käsitteistö koskien tekoälyä ja sen hyödyntämistä.

Data-alustat ja integraatiot	Tekoälyn hyödyntäminen perustuu useimmiten datan hallintaan. Organisaatiolla on suotavaa olla tietokantojen, tietovarastojen tai tietoaltaiden käsittelyn osaamista sekä datan siirtämiseen liittyvää integraatioiden osaamista.
Ohjelmistokehitys	Dataa ja tekoälyä hyödyntävä ratkaisu toteutetaan usein ohjelmoimalla, jolloin vakiintuneet ohjelmistojen versionhallinnan ja laadunhallinnan periaatteet ovat edellytys myös tekoälyhankkeiden onnistumiselle.
Juridinen osaaminen	Organisaatiolla on oltava tarvittava juridinen osaaminen koskien tietosuoja-asetusta ja datanhallintaa, mutta myös tekoälyä ja automaattista päätöksentekoa koskeva lainsäädäntö on hallittava.
Eettinen osaaminen	Organisaatiolla on hyvä olla teknologian etiikkaan liittyvää osaamista. Organisaatioilla on suotavaa olla datan ja tekoälyn hyödyntämisen eettinen ohje. Eettinen ohjaa kuvaa lyhyesti organisaation periaatteet koskien datan ja tekoälyn hyödyntämistä.

1.4 Tekoäly, koneoppiminen ja yleisimmät haasteet

Tekoälyratkaisun älykkyys useimmiten opitaan datasta koneoppimisen avulla. Ratkaistavasta ongelmasta riippuu, mikä koneoppimisen muodoista sopii parhaiten ratkaisun pohjaksi. Erilaisissa koneoppimisen muodoissa on omat vahvuutensa, mutta myös heikkoutensa. Oheisessa taulukossa on lyhyesti kuvattu tyypillisimmät koneoppimisen muodot, niiden toimintaperiaate sekä merkittävimmät haasteet turvallisen kehittämisen kannalta.

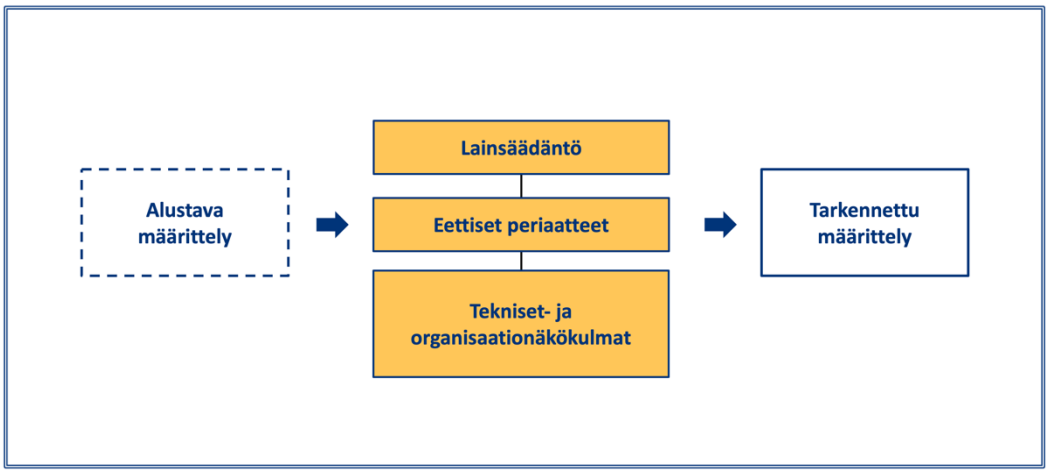
Yleisin koneoppimisen muoto on ohjattu oppiminen, jossa tekoälylle annetaan opetusdatan muodossa esimerkkejä syötteistä sekä ihmisten tuottamia haluttuja tuloksia. Tässä asetelmassa oppimisen onnistuminen riippuu lähes täysin siitä, mitä opetusaineistona annettu esimerkkidata sisältää (1).

Taulukko 3: Koneoppimisen muodot ja niiden yleisimmät haasteet turvallisen kehittämisen kannalta.

Oppimisen muoto	Periaate, esimerkkejä ja yleisimmät haasteet
Ohjattu oppiminen	Tekoälylle annetaan esimerkkejä sekä syötteestä että halutusta lopputuloksesta. Tekoälyn tehtävä on oppia ennustamaan uusille syötteille paras mahdollinen lopputulos. Esimerkkejä ohjatusta oppimisesta ovat automaattinen päätöksenteko, suositukset, kuvantunnistus, luokittelut, kategorisoinnit, sekä jatkuva-arvoisen muuttujan (esim. tulot, kesto) ennustaminen. Merkittävimmät haasteet liittyvät siihen, ovatko tekoälyn opetusdatasta oppimat ja tuottamat vastaukset reiluja vai esiintyykö niissä mahdollisesti syrjintää (1).

Ohjaamaton oppiminen	Tekoälyn opettamisessa ei ole käytettävissä oikeita vastauksia. Tekoälyn on opittava esimerkiksi ryhmittelemään aineiston näytteet samankaltaisuuden mukaan. Esimerkkejä ohjaamattomasta oppimisesta ovat asiakasanalyysit, segmentoinnit, klusterointi (ryhmittely) ja moniulotteisin datan dimensionaalisuuden pienentäminen. Merkittävimmät haasteet liittyvät esimerkiksi henkilöistä kerättyjen aineistojen analyyseihin: riittävän tarkan analyysin saavuttaminen voi edellyttää laajahkoja tietoaineistoja. Lisäksi aineistot voivat todellisuudessa kuvata todellista ilmiötä hyvinkin puutteellisesti, vaikeuttaen tulosten tulkintaa.
Vahvistusoppiminen	Tekoäly, toimii itsenäisenä agenttina, joka opetetaan ennustamaan optimaalinen toiminta agentin havaitsemassa tilanteessa. Esimerkki mahdollisesta vahvistusoppimisen soveltamisesta on autonomisen auton ohjaus kamera- ja tutkainformaation perusteella. Pitkälle kehittyneet kielimallit voivat hyödyntää vahvistusoppimista sanojen generointiin tekstimuotoisessa kontekstissa. Merkittävimmät haasteet turvallisen kehittämisen kannalta liittyvät agentin opettamisessa tarvittavan palkinnon määrittämiseen; mitä opettamisessa pidetään onnistumisena.

2 Tekoälyratkaisujen määrittely



Kuva 2: Määrittelyn vaiheet.

Datan ja tekoälyn hyödyntämistä tavoittelevien hankkeiden ensimmäinen vaihe on määrittely. Aluksi määritellään järjestelmän käyttötarkoitus, tavoitteet ja teknisen ratkaisun pääpiirteet. Tämän jälkeen oleellinen toteutuksen kannalta relevantti lainsäädäntö voidaan tunnistaa ja huomioida ratkaisun tarkemmassa suunnittelussa. Määrittelyvaiheessa on tärkeää kuvata myös ratkaisun kannalta tärkeimmät eettiset periaatteet ja näkökulmat. Määrittelyn kannalta oleellista on tunnistaa, minkälaiselle teknologiselle pohjalle ja asiantuntijaosaamiselle hanke perustuu. Juridisten, eettisten ja toteutusteknisen näkökulmien huomioiminen mahdollistaa alkuperäisen määrittelyn

tarkentamisen riittävälle tarkkuustasolle. Ratkaisua kehittävään organisaatioon voi liittyä erityislainsäädäntöä, jonka vaatimukset on erikseen huomioitava.

2.1 Lainsäädännön huomioiminen tekoälyhankkeen määrittelyssä

Tärkeimmät dataa ja tekoälyn hyödyntämistä säätelevät lait ovat tietosuoja-asetus (4), sekä automaattista päätöksentekoa koskevat lait (7), (8), sekä tekoälyasetus (5). Lainsäädännön tarkoituksena on suojella kansalaisten perusoikeuksien toteutumista. Tavoitteena on taata kansalaisten yksityisyyden suoja henkilötietojen käsittelyn yhteydessä sekä varmistaa, että mahdolliset päätökset ovat oikeudenmukaisia. Alla kuvatussa taulukossa on kuvattu tärkeimmät käytettävien tietolähteiden määrittelyä koskevat ohjeet.

Taulukko 4: Ohjeita tekoälyjärjestelmien hyödyntämien tietoaineistojen määrittelyyn.

Aihe	Tietolähteiden määrittelyä koskevat ohjeet
Tietolähteiden kuvaaminen	Kuvaa määrittelyssä, hyödynnetäänkö organisaation omia tietovarantoja, tietoluvan edellyttämiä ulkoisia aineistoja vai esimerkiksi avoimia tietolähteitä.
Henkilötiedot, tietosuoja-asetus	Kuvaa määrittelyssä, hyödynnetäänkö järjestelmässä henkilötietoja. Kuvaa tietolähteet ja tietosisällöt. Kuvaa määrittelyssä, onko henkilötietojen käyttämiseen olemassa rekisteröityjen suostumus. Määrittelyn tulee kuvata tietosuoja-asetuksen minimiperiaatteiden sekä oletusarvoisen ja sisäänrakennetun tietosuojan toteuttaminen. Huomioi, että henkilötietoja käytettäessä määrittelyvaiheessa on toteutettava tietosuoja-asetuksessa vaadittava vaikutusten arviointi riskiarvioineen.
Tietojen elinkaari	Kuvaa määrittelyssä, onko tietojen hyödyntäminen kerta- vai jatkuvaluonteista, ja miten tietojen elinkaarenhallinta toteutetaan.
Henkilötietoja sisältävän rekisterin syntyminen	Tietojen kerääminen ja käsittely saattavat johtaa uuden henkilörekisterin syntymiseen. Määrittelyn on otettava kantaa mahdollisen henkilörekisterin syntymiseen ja siihen liittyviin vastuisiin.
Oikeusperuste	Organisaatio voi hyödyntää dataa, henkilötietoja ja uusia teknologioita lakisääteistä tehtävää suorittaessaan. Määrittelyvaiheessa on pyrittävä tunnistamaan lainkohdat, jonka perusteella tietoja käytetään ja tekoälyratkaisua kehitetään. Jos käsittely perustuu rekisteröityjen suostumukseen, määrittelyssä on varmistuttava, että käyttötarkoitus vastaa suostumuksen sisältöä.

Tekoälyä hyödyntävien järjestelmien lainmukaisuus ei liity pelkästään käytettyihin tietoaineistoihin, vaan myös vahvasti tekoälyratkaisujen käyttötarkoituksiin. Alla olevassa taulukossa on kuvattu ohjeita koskien tekoälyjärjestelmän käyttötarkoituksen määrittelemistä.

Taulukko 5: Ohjeita tekoälyjärjestelmän käyttötarkoituksen määrittelyyn.

Aihe	Tietojen käyttötarkoituksen määrittelyä koskevat ohjeet
Millä tasolla tietoja hyödynnetään?	Jos tavoitteena on analysoida tiedoista ilmiöitä tai ymmärtää asiakkaita yleisesti, kyseessä on strateginen käyttötarkoitus. Tässä tapauksessa tuotetaan aggregoitua tai tilastollista tietoa, jota ei jälkikäteen voida yhdistää yksittäisiin henkilöihin. Jos tavoitteena on hyödyntää henkilötietoja esim. asiakasmassaa koskevien käsittelyprosessien, asianhallinnan tai vastuiden sujuvoittamiseksi, kyseessä voi olla taktinen käyttötarkoitus. Tällöin järjestelmässä ei tehdä päätöksiä yksittäisistä asioista, vaan pyritään sujuvoittamaan prosesseja. Jos tavoitteena on hyödyntää tietoja yksittäisen henkilön tilanteen käsittelyssä tai päätöksenteossa, kyseessä on operatiivinen käyttötarkoitus. Operatiivisessa päätöksenteossa ihmisen ja järjestelmän roolit on kuvattava mahdollisimman tarkasti. Määrittelyssä on kuvattava, millä tasolla henkilötietoja aiotaan hyödyntää.
Automaattinen päätöksenteko	Määrittelyssä on kuvattava, mikäli tavoitteena on täysin automaattinen päätöksenteko, ja mikä on mahdollinen ihmisen rooli päätöksen syntymisessä. Määrittelyssä on kuvattava, mikäli tekninen ratkaisu toteuttaa ns. profilointia, eli henkilöiden henkilökohtaisten ominaisuuksien arviointia. Profiloinnin yhteys päätöksentekoon on kuvattava määrittelyssä. Huomioi automaattista päätöksentekoa julkishallinnossa koskeva lainsäädäntö (7), (8).
Tekoälyasetus	Huomio määrittelyssä, että EU:ssa valmistellaan tekoälyasetusta, joka asettaa vaatimuksia tekoälyjärjestelmien toteuttamiselle ja käytölle (5). Määrittelyssä on suotavaa ottaa kantaa, kuuluuko kehitettävä järjestelmä suuren, vähäisen vai minimaalisen riskin järjestelmiin. Määrittelyssä on huomioitava riskitasoa vastaavien prosessien toteuttaminen.

2.2 Eettisten näkökulmien huomioiminen hankkeen määrittelyssä

Ajantasaisen ja suunnitteilla olevan lainsäädännön noudattaminen dataa ja tekoälyä hyödyntävissä hankkeissa ei riitä. Teknologioiden kehittyessä lainsäädäntö ei useinkaan kykene uudistumaan riittävän nopeasti. Tästä syystä kehittyneiden teknologioiden hyödyntämisessä on pohdittava teknisten ratkaisujen eettisyyttä ja pitkän aikavälin vaikutuksia yhteiskuntaan. Kattava ohjeistus eettisten näkökulmien huomioimiseen löytyy EU:n luotettavan tekoälyn eettisestä ohjeistuksesta (6).

Taulukko 6: Ohjeita luotettavan tekoälyn toteuttamiseen.

Aihe	Ohjeet luotettavan tekoälyn toteuttamiseen
Luotettavan tekoälyn periaatteiden toteutuminen	Määrittelyn tulisi ottaa kantaa, miten seuraavat periaatteet toteutuvat suunnitteilla olevassa järjestelmässä: Itsemääräämisoikeuden kunnioittaminen: järjestelmä ei saa manipuloida, johtaa harhaan eikä holhota ihmisiä. Vahinkojen välttäminen: järjestelmä ei saa aiheuttaa vahinkoa. Tämä sisältää sekä henkisen että fyysisen koskemattomuuden suojelun. Oikeudenmukaisuus: aineellinen näkökulma edellyttää ihmisten tasapuolista kohtelua sekä hyötyjen tasaista jakautumista yhteiskunnassa. Menettelyllinen näkökulma edellyttää, että tekoälyn tekemiä päätöksiä tai ratkaisuja voidaan riitauttaa ja niihin voidaan hakea muutosta. Selitettävyys: tekoälyn tekemien päätösten ja ratkaisujen tulisi olla selitettäviä, mutta myös kehittämisen prosessin tulisi olla läpinäkyvä ja avoin.
Luotettavan tekoälyn vaatimusten täyttyminen	Määrittelyssä on suotavaa kuvata, miten seuraavat vaatimukset toteutuvat suunnitellussa järjestelmässä: • Perusoikeudet, ihmisen toimijuus ja ihmisen suorittama valvonta • Tekninen luotettavuus ja turvallisuus • Yksityisyyden suoja ja datan hallinta • Avoimuus • Monimuotoisuus, syrjimättömyys, oikeudenmukaisuus • Yhteiskunnallinen ja ekologinen hyvinvointi • Vastuuvelvollisuus

2.3 Teknisten reunaehtojen ja asiantuntijaosaamisten määrittely

Määrittelyn on kuvattava pääpiirteet teknisen ratkaisun toteuttamiseksi tai hankkimiseksi. Määrittelyssä on myös tarpeen kuvata, millä tavalla hankkeen vaatimat asiantuntijaosaamiset aiotaan hankkia. Alla olevassa taulukossa on esitetty määrittelyä helpottavia kysymyksiä sekä ohjeita määrittelyn toteuttamiseksi.

Taulukko 7: Ohjeita teknisten reunaehtojen ja asiantuntijaresurssien määrittelyyn.

Aihe	Ohje määrittelyä varten
Hyödyntääkö tekninen ratkaisu kaupallisia valmistuotteita?	Jos ratkaisu hyödyntää valmisratkaisuja, on määrittelyä varten selvitettävä vastuun jakautuminen valmisratkaisun toteuttajan ja hankkivan osapuolen välille.
Hyödyntääkö tekninen ratkaisu avoimen lähdekoodin ratkaisuja?	Määrittelyssä on kuvattava toteuttavalle osapuolelle jäävät vastuut. Määrittelyssä on myös tuotava esiin suunnitelma, miten järjestelmän tuki, ylläpito, ja virheiden korjaaminen aiotaan järjestää.
Hyödyntääkö tekninen ratkaisu valmiiksi opetettuja malleja?	Määrittelyssä on kuvattava, miten mahdolliset virheet tai puutteet valmiiksi opetetussa mallissa käsitellään, jos toteuttajalla on rajalliset mahdollisuudet muokata ja käsitellä valmiiksi opetettua mallia.

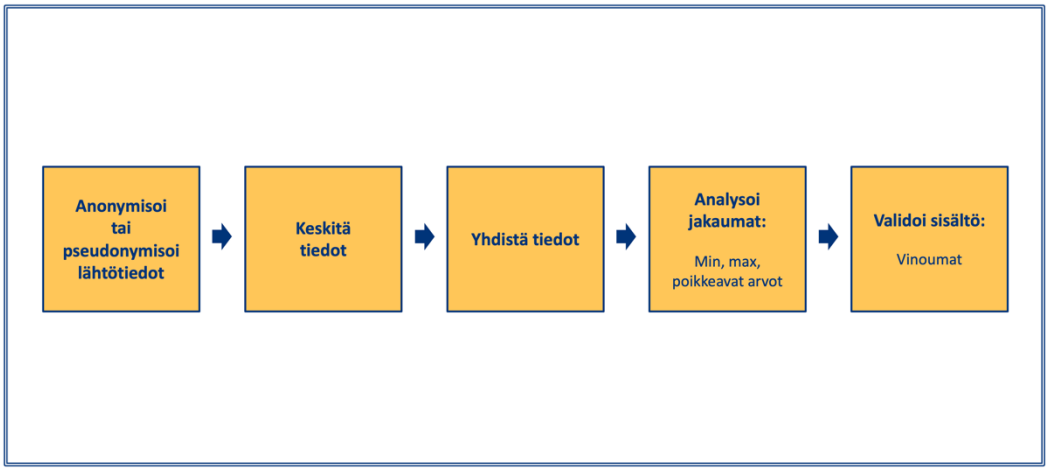
Missä teknisessä ympäristössä tekninen ratkaisu kehitetään ja ylläpidetään?	Määrittelyssä on kuvattava, miltä osin järjestelmä toimii organisaation omilla fyysisillä palvelimilla ja miltä osin esim. pilvessä. Määrittelyssä on myös kuvattava, missä ratkaisun käsittelemä data sijaitsee, ja siirretäänkö dataa pilveen ja EU alueen ulkopuolelle. Lisäohjeita löytyy Virta-arkkitehtuurista (3).
Toteutetaanko tekoälyratkaisu organisaation omalla henkilöstöllä?	Jos hanke toteutetaan organisaation omalla henkilökunnalla, on varmistettava riittävä ja sopiva osaamisen taso ja monipuolisuus, sekä riittävät organisaation prosessit, joilla jatkuva kehittäminen ja ratkaisun elinkaaren hallinta toteutetaan. Määrittelyssä on suotavaa kuvata, miten osaamisen jatkuvuus turvataan.
Hyödynnetäänkö hankkeessa organisaation ulkopuolisia asiantuntijoita?	Ulkopuolisia asiantuntijoita hankittaessa on huomioitava luotettavuus, salassapitosopimukset sekä tietosuoja- ja tietoturvakäytännöt. Ulkopuolisille asiantuntijoille on huolehdittava kehitystyössä riittävät käyttöoikeudet, turvalliset yhteydet sekä autentikointi. Määrittelyssä on suotavaa kuvata sidosryhmät ja niiden vastuut.

Määrittelyvaihe saattaa sisältää virallisia päätöksiä koskien hankkeen toteuttamista. Päätöksissä arvioidaan ratkaisun tavoitetta, eettisyyttä, lainmukaisuutta, ja käytettävien tietojen oikeasuhtaisuutta. Päätökset perusteluineen on tallennettava huolellisesti myöhempää mahdollista käyttöä varten

3 Datan keruu, tallennus ja käsittely

3.1 Datan käsittelyn vaiheet ja avainkäsitteet

Datan keruu-, tallennus- ja käsittelyvaiheessa siirrytään suunnitelmasta toimintaan. Data- ja tekoälykehittämiseen liittyy väistämättä toimivaan ratkaisuun tähtäävä kokeileminen. Hankkeissa on kuitenkin varmistuttava, että alkuperäinen käyttötarkoitus, toimintaperiaate ja lainmukaisuus eivät muutu kokeilujen seurauksena. Alla olevassa kuvassa on esitetty datan keräämisen ja käsittelyn tyypilliset vaiheet.



Kuva 3: Datan käsittelyn vaiheet.

Turvallisen kehittämisen kannalta kriittiset toimenpiteet liittyvät aineistoihin, jotka sisältävät henkilötietoja. Tällöin oleellista on tunnistaa kerättyjen tietojen alkuperäinen käyttötarkoitus, ja mahdolliseen uuteen käyttötarkoitukseen liittyvät tietolupaprosessit ja mahdollisten rekisteröityjen suostumusten kerääminen. Toiminnan ohjeistus riippuu hyvin paljon tietojen käyttötarkoituksesta. Lisätietoa turvallisen arkkitehtuurin suunnittelusta ja toteutuksesta löytyy Virta-arkkitehtuurista (3).

Henkilötietoja sisältävien data-aineistojen käsittelyssä on välttämätöntä tuntea seuraavat avainkäsitteet:

- Anonymisointi: henkilötietojen suorien tunnisteiden (nimi, henkilötunnus, osoite, jne.) salaaminen siten, että rekisteröidyn identiteetti ei ole jälkikäteen kohtuullisilla resursseilla palautettavissa.
- Pseudonymisointi: henkilötietoaineiston suorien tunnisteiden (nimi, henkilötunnus, osoite, jne.) salaaminen siten, että rekisteröidyn identiteetti ei paljastu. Identiteetti on kuitenkin jälkikäteen palautettavissa.

3.2 Datan keruu, anonymisointi ja yhdistäminen eri käyttötarkoituksissa

Kolme ensimmäistä datan käsittelyn vaihetta liittyvät usein tiiviisti toisiinsa. Näiden vaiheiden toteutus riippuu myös kehitettävän järjestelmän käyttötarkoituksesta. Seuraavaksi esitetään tyypilliset käyttötarkoitukset sekä niihin liittyvät tietojen käsittelyn ja yhdistämisen ohjeet.

Alla olevassa taulukossa on käsitelty dataa ja tekoälyä hyödyntävän järjestelmän toteutusta, jonka käyttötarkoituksena on strategisen tason tiedon hyödyntäminen, esimerkiksi johtamisen tueksi.

Taulukko 8: Tietojen käsittelyn ohjeet strategiseen tiedon hyödyntämiseen.

Käyttötarkoitus	Tietojen käsittelyn ja yhdistämisen ohjeet
Tietoja käsitellään strategisella tasolla johtamisen tukena, tilastollisen aineiston tavoin.	Jos aineisto on ulkoisen rekisterinpitäjän hallinnassa, tietoaineistoon on haettava lupa tietolupaprosessien mukaisesti. Jos aineistossa yhdistetään eri rekisterinpitäjien aineistoja, on järjestettävä tietoturvallinen ympäristö, jossa käsittely tapahtuu. Tunnistetiedoille luodaan pseudonymisoidut tunnisteet salausavaimen avulla, salausavain tallennetaan turvallisesti. Jokaisen rekisterin suorat tunnisteet korvataan pseudonymisoiduilla tunnisteilla lähtöjärjestelmässä/rekisterissä ennen yhdistämistä. Pseudonymisoidut aineistot keskitetään. Aineistojen yhdistäminen tapahtuu ilman suoria tunnisteita, pseudonymisoitujen tunnisteiden avulla. Yhdistämisen jälkeen pseudonymisoidut tunnisteet voidaan poistaa. Aineistoa voidaan käyttää tilastollisen mallin muodostamiseen. Jos käsittely on tapahtunut erillisessä tietoturvallisessa ympäristössä, tilastolliset tulokset tuodaan järjestelmästä sovittuun käyttötarkoitukseen.

Dataa ja tekoälyä hyödyntävä järjestelmä voi pyrkiä tehostamaan organisaation sisäisiä prosesseja. Tässä tarkoituksessa ei yleensä ratkaista yksittäisten henkilöiden asioita. Alla olevassa taulukossa on kuvattu ohjeet tietojen käsittelyyn taktisessa käyttötarkoituksessa.

Taulukko 9: Tietojen käsittelyn ohjeet taktisen tason tietojen hyödyntämiseen.

Käyttötarkoitus	Tietojen käsittelyn ja yhdistämisen ohjeet
Tietojen käyttäminen on taktisen tason hyödyntämistä, kuten esimerkiksi käsittelyprosessin sujuvoittamista.	Varmista, että organisaatiota koskeva lainsäädäntö mahdollistaa henkilötietojen käyttämisen toiminnan kehittämisessä, työn tehostamisessa ja työn ohjauksessa. Varmista, että toteutustyöt noudattavat määrittelyvaiheen reunaehtoja datankäsittelyn kaikissa vaiheissa Pseudonymisoi tai anonymisoi kehittämisessä tarvittavat tiedon mahdollisimman aikaisin (sisään rakennettu tietosuoja) Huomioi, että käsiteltävien tietojen laajuuden on oltava perusteltavissa olevassa suhteessa käyttötarkoitukseen (minimiperiaate). Huolehdi, että tietojen hyödyntäminen noudattaa organisaatiosi käyttöoikeuksien sovittuja periaatteita.

Alla olevassa taulukossa on puolestaan kuvattu tietojen käsittelyn ohjeet operatiivisella tasolla toimivalle dataa ja tekoälyä hyödyntävälle järjestelmälle. Tietojen hyödyntäminen operatiivisella tasolla vaatii käsittelyn lainmukaisuuden varmistamista, koska ratkaisun toiminnalla on huomattava vaikutus henkilön asioiden käsittelyyn.

Taulukko 10: Tietojen käsittelyn ohjeet operatiivisella tasolla toimivalle järjestelmälle.

Käyttötarkoitus	Tietojen käsittelyn ja yhdistämisen ohjeet
Henkilötietoa käytetään yksilöllisessä palvelemisessa tai päätöksenteossa. Henkilötietojen käsittelyllä on suora vaikutus rekisteröityyn.	Ratkaisun kehittäminen ja hyödyntäminen tapahtuu anonymisoidulla tai pseudonymisoidulla aineistolla, mikäli mahdollista. Henkilöä koskevan päätöksenteon yhteydessä on käytettävä vain oikeaa, laadukasta ja luotettavaa tietoa. Tiedon laatu on kyettävä validoimaan. Operatiivisissa päätöksentekotilanteissa pääsy tietoihin on vain määritetyillä henkilöillä. Tiedon siirtäminen järjestelmien välillä on oltava turvallista. Tarvittaessa järjestelmän on kyettävä tallentamaan rekisteröidyn suostumus tietojen käyttöön mainitussa käyttötarkoituksessa. Rekisteröidyn tiedot ja tiedon avulla tuotetut ennusteet ja rikasteet on kyettävä erottamaan toisistaan tiedon tallennuksessa. Jos kyse on digitaalisesta palvelusta, tietosuojaselosteen on kuvattava tietojen käyttötarkoitus ja laajuus. Datan käsittelyssä ei tule käyttää syrjintäkriteereiksi määriteltyjä tietoja.

Aiemmat käyttötarkoituksen olettivat, että tietoaineisto on esimerkiksi primäärien järjestelmien keräämänä olemassa. Usein tarvittava tietoaineisto voidaan joutua keräämään sovellusten tai kyselyiden avulla. Alla olevassa taulukossa on esitetty tietojen käsittelyn ohjeet kyselyn avulla kerättävän aineiston käsittelyyn.

Taulukko 11: Tietojen käsittelyn ohjeet uuden aineiston keräämiseen.

Käyttötarkoitus	Datan käsittelyn ja yhdistämisen ohjeet
Data-aineistoa ei lähtötilanteessa ole. Data-aineisto muodostetaan esim. kyselyllä.	Kysymysten tulee huomioida kaikkien erilaisten vastaajien taustat. Kysymysten muotoilun tulee mahdollistaa hyvin erilaisten vastausten antamisen ilman, että kysely ohjaa vastaajaa. Väärät olettamat vastaajista eivät saa rajoittaa kysymyksiin vastaamista. Monivalintavastaukset tulee sanoittaa huolellisesti siten, että vaihtoehdot on mahdollista ymmärtää johdonmukaisesti samalla tavalla. Mahdollisissa vapaatekstikentissä on huomioitava henkilötasoisten tunnisteiden mahdollinen syöttäminen, ja niiden esiintyessä henkilötietojen anonymisointi. Aineistoa kerättäessä on tallennettava rekisteröidyn suostumus tietojen tallentamiseen ja hyödyntämiseen käyttötarkoituksen mukaisesti. Kerättyä aineistoa ei saa myöhemmin käyttää uusiin käyttötarkoituksiin ilman suostumusta.

3.3 Numeerisen datan alustava analyysi

Kun käytettävissä oleva aineisto on kerätty yhteen paikkaan, yleensä toteutetaan datan muuttujakohtainen analyysi. Sen avulla muodostetaan kokonaiskuva aineistosta. Usein lasketaan esimerkiksi numeeristen muuttujien keskiarvot ja hajonta, tai pyritään tunnistamaan ilmeiset mittausvirheet. Tavoitteena on tunnistaa ilmeiset puutteet tiedon keruussa ja yhdistämisessä sekä välttää sellaisten siirtyminen mallinnukseen.

3.4 Datan sisällön validointi ja vinoumat

Tekoälyn turvallisen hyödyntämisen yksi tunnetuimpia haasteita ovat data-aineistojen vinoumat ja niiden vaikutus lopputuloksena syntyneeseen älykkääseen järjestelmään. Vinoumien läsnäolo yleisesti tarkoittaa sitä, data ei vastaa kaikilta osin todellisuutta, vaikka usein näin tulkitaan. Tämän seurauksena tekoälymalli ei toimi toivotusti tai suunnitellusti kaikissa tilanteissa. Sen sijaan tekoälymallin toiminnassa on ei-toivottuja piirteitä, joiden eliminointi voi olla vaikeaa.

Vinoumia esiintyy datan käsittelyn, mallin opettamisen ja mallin käytön vaiheissa. Alla olevassa taulukossa on kuvattu ohjeita erilaisten vinoumien käsittelyyn. Lisätietoa datan vinoumiin liittyen löytyy mm. syrjimättömän tekoälyn arviointikehikosta (1).

Taulukko 12: Datan käsittelyvaiheen vinoumia ja ohjeita niiden käsittelyyn.

Vinouma	Ohjeita vinoumien käsittelyyn
Edustavuusvinouma	Ongelmana on, että data-aineisto tai kyselyn tulokset eivät edusta koko kohderyhmää. Tällöin voidaan selvittää datassa olevien demografisten tai vastaavien taustamuuttujien avulla, ovatko kaikki tavoitellut kohderyhmät edustettuja aineistossa. Jos aineisto ei ole riittävän kattava, kerää lisää aineistoa. Jos mahdollista, testaa kyselyä tai kehitä kysely hyvin monipuolisten ja taustoiltaan erilaisten kohdejoukkojen edustajien avustuksella. Mahdollisuuksien mukaan tulee selvittää erilaiset kanavat ja työkalut eri kohderyhmien tavoittamiseksi.
Otantavinouma	Otantavinouman aiheuttaa aineiston valikoituminen. Esimerkiksi kyselyyn vastaaminen, eli näytteistys ei ole satunnaista. Voi olla, että tietyt ominaisuudet omaavat henkilöt vastaavat kyselyyn herkemmin kuin toiset, vääristäen aineiston muodostumista. Pyri vaikuttamaan valikoitumiseen siten, että taustoiltaan ja tilanteeltaan erilaiset kohdejoukot tulevat edustetuiksi.
Mittausvinouma	Datassa esiintyvä muuttujan arvo voi olla virheellinen tai tulkinnanvarainen (subjektiivinen). Esimerkiksi kyselytutkimuksen kysymykset ja vastausvaihtoehdot tulee suunnitella siten, että väärinymmärrysten ja tulkintavirheiden mahdollisuus minimoituu. Kerätystä aineistosta analysoidaan jokaisen muuttujan arvojen jakauma. Selvästi poikkeavat arvot tulee analysoida tarkasti, ja mahdollisesti poistaa analyysistä.
Luokitteluvinouma	Tavoitteena on hyödyntää ohjattua koneoppimista, jossa tekoälymallille annetaan syötedatan lisäksi ihmisen luokittelemia ns. oikeita vastauksia. Luokitteluvinouma johtuu ihmisten tekemien luokitteluiden epäjohdonmukaisuudesta. Ohjeistus eri luokkien käyttämisestä tulee olla johdonmukainen ja riittävän tarkka. Lisäksi mahdolliset virheelliset luokittelut tulee pyrkiä tunnistamaan ja mahdollisesti poistamaan aineistosta. Virheellisten luokitusten tunnistaminen saattaa edellyttää mallin opettamista, tulosten ja ennustevirheiden syvällistä analyysiä. Poikkeavien luokitusten poistamiselle tulisi löytää järkevä selitys.

Puuttuvan muuttujan vinouma	Asetelmassa tavoitellaan ennustemallin tekemistä, mutta tiedetään oleellisen muuttujan puuttuvan mallista. Tällaisessa tilanteessa etsitään tietolähde, josta tieto voisi olla saatavilla. Kannattaa myös pohtia, onko puuttuvaa tietoa mahdollista approksimoida tai ennustaa välillisesti toisten muuttujien avulla. Mahdollista välillistä syrjintää on kuitenkin vältettävä.
Interaktiovinouma	Datan luullaan kertovan käyttäjän ominaisuuksista, vaikka data syntyy käyttäjän ja tietojärjestelmän vuorovaikutuksesta, jossa järjestelmällä on ohjaileva tai toimintaa rajaava vaikutus. Järjestelmän ohjaava vaikutus aineiston sisältöön on tiedostettava ja huomioitava mallinnuksessa, mikäli mahdollista.
Loukkaava sisältö	Opetusdata sisältää eksplisiittisesti loukkaavaa sisältöä. Tällainen sisältö tulee poistaa ennen mallinnusta.

3.5 Tekstipohjaisten aineistojen käsittely

Tekoälyn hyödyntämisessä nopeasti yleistyvä käyttötapaus on vapaamuotoisen, eirakenteellisen tekstiaineiston hyödyntäminen tekoälyn opettamisessa. Aineiston lähteenä voi olla esimerkiksi avoimet tekstimassat (wikipedia), organisaation keräämät ja ylläpitämät tekstiaineistot (asiakaspalautteet, potilaskertomukset), tai sovellusta varten erikseen kerättävät tai muodostettavat tekstiaineistot (chatbot -opetusdatat).

Tekstiä hyödyntävät tekoälyratkaisut voivat esim. analysoida tekstimassojen sisältöä, tai toteuttaa keskustelevia käyttöliittymiä (chatbotit). Turvallisen kehittämisen kannalta tekstiaineistojen hyödyntämistä koskevat samat lainalaisuudet kuin kategorisen tai numeerisen tiedon tapauksessa; henkilötietojen käsittelyn on oltava lainmukaista, syrjimätöntä ja erityisesti yksityisyyden suojasta on huolehdittava kaikissa käsittelyn vaiheissa.

Taulukko 13: Ohjeita tekstiaineistojen käsittelyyn.

Käyttötapaus	Ohjeita tekstiaineiston käsittelyyn
Käytettävissä oleva tekstiaineisto voi sisältää henkilötietoja, kuten esimerkiksi etunimi, sukunimi, osoite, sähköpostiosoite, postiosoite	Tekstiaineisto on anonymisoitava. Etunimien ja sukunimien tunnistamiseen on mahdollista käyttää etunimi- ja sukunimitilastoja. Sähköpostiosoitteiden tunnistamiseen on mahdollista käyttää säännöllisiä lausekkeita (regular expression), jonka avulla @ merkin ja domainin loppuosan sisältävät merkkijonot tunnistetaan. On myös mahdollista opettaa mallipohjaisia ratkaisuja vastaaviin käyttötapauksiin, jotka sallivat kirjoitusvirheitä. Sanaluokkien tunnistamiseen on saatavilla kielioppiin ja koneoppimiseen perustuvia malleja. Niiden avulla on mahdollista tunnistaa ja anonymisoida henkilötietoja. Mallipohjaisen anonymisoinnin kehittämisessä voidaan käyttää myös synteettistä, keinotekoista aineistoa.
Tavoitteena on toteuttaa chatbot ja tuotetaan esimerkkilauseita ja vastauksia asiantuntijatyönä	Varmistetaan, että chatbotin käyttäjien syötteitä simuloiva esimerkkiaineisto on sisällöllisesti kattava ja erilaisia kieliasuja saman asian ilmeisemiseksi on riittävästi. Tämä vähentää chabotin tekemiä virheellisiä syötteen luokitteluita, joka pääsääntöisesti johtaa arvaamattomiin seurauksiin. Varmistetaan, että chatbotin tuottamat vastaukset erilaisissa dialogin tilanteissa eivät ole syrjiviä tai loukkaavia.

Chatbotin tai sovelluksen vapaatekstinkentän käytön yhteydessä syötettävät mahdolliset henkilötiedot ja niiden käsittely	Käyttäjän syöttämien tekstien tallentamisessa on huomioitava, että käyttäjä voi ohjeistuksen vastaisesti syöttää henkilötietoja järjestelmään. Järjestelmän on kyettävä anonymisoimaan henkilötietoja sisältävä syöte ennen tietojen lokitusta ja hyödyntämistä chatbotin keskusteluominaisuuksien kehittämiseen.

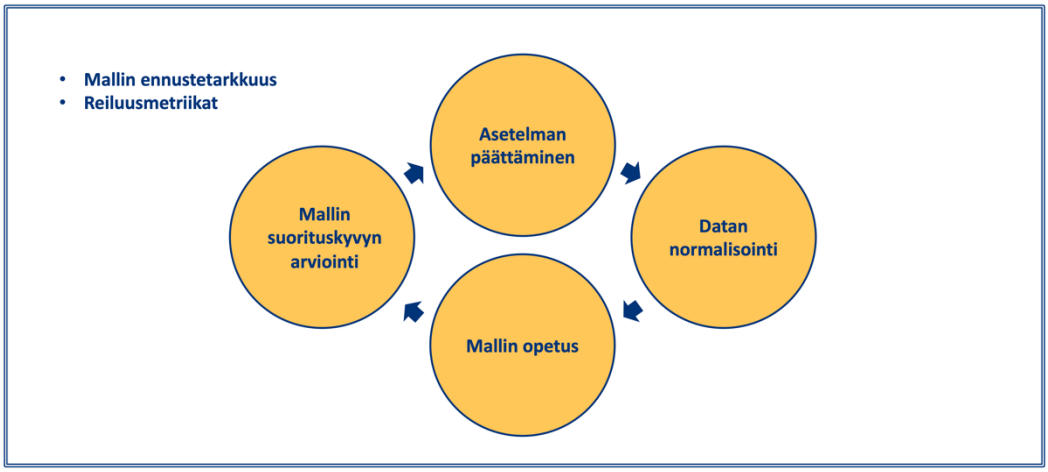
4 Tekoälyn opettaminen

Ennen tekoälyn opettamista tulee suunnitella kehittämisessä käytettävä prosessi ja niissä tarvittavat järjestelmät. Tekoälyn opettaminen voi tapahtua joko kokonaan mallia kehittävän organisaation infrastruktuurissa, jolloin malli ja mallin opettamisessa tarvittavat tiedot ovat ainoastaan asianmukaisten henkilöiden saavutettavissa. Joissain tilanteissa tekoälyn kehittämiseen liittyvä ohjelmointi kannattaa toteuttaa kehittäjän omalla laitteistolla, ja viedä kehittämisessä syntyvä lähdekoodi versionhallintaan. Versionhallinnasta julkaistaan uusimmat versiot tuotantoon. Kuhunkin tilanteeseen sopiva lähestymistapa riippuu opettamisessa tarvittavista tiedoista sekä mallin käyttötarkoituksesta. Alla olevassa taulukossa on kuvattu kaksi erilaista lähestymistapaa ohjeineen.

Taulukko 14: Ohjeita erilaisiin kehitysprosesseihin.

Kehitysprosessin arkkitehtuuri	Ohjeet kehittämiseen
Ympäristöt ja kehittäminen tapahtuvat ainoastaan organisaation järjestelmissä.	Huolehdi, että ratkaisun tuotantoversion ja kehitysversioiden tarpeisiin sopivat tietokannat ja datan tallennuspaikat ovat tietoturvallisia, ja niihin on pääsy ainoastaan asianmukaisilla henkilöillä. Huolehdi, että tiedonsiirto tietojen ja ratkaisun välillä on turvallista.
Kehittäminen tapahtuu kehittäjän omalla laitteella, josta lähdekoodi viedään versionhallinnan kautta tuotantoon.	Huolehdi, että tiedonsiirto kehittäjän laitteen ja organisaation muun infrastruktuurin välillä on turvallinen, ja pääsy järjestelmiin on ainoastaan asianmukaisilla henkilöillä. Huolehdi, että kehittäjän laitteistolle ei missään olosuhteissa tallennu henkilötietoa.

Itse mallin opettamisessa on tärkeää noudattaa yleisiä koneoppimisen kehittämisen käytäntöjä. Näin varmistetaan mallin toimivuus sille suunnitellussa käyttötarkoituksessa. Mallinnusvaiheen tyypilliset työvaiheet ovat mallinnusasetelman määrittäminen, datan esikäsittely, mallin opettaminen, mallin analysointi. Tämän jälkeen mallia ja asetelmaa muokataan, kunnes malli toimii odotusten mukaisesti.



Kuva 4: Tekoälymallin opettamisen vaiheet.

Alla olevassa taulukossa on kuvattu mallin opettamisen vaiheet, sekä ohjeita vaiheiden toteuttamiseen. Mallin reiluutta ja mahdollista oikaisua koskevia ohjeita on kuvattu mm. syrjimättömän tekoälyn arviointikehikossa (1).

Taulukko 15: Ohjeita mallin opettamisen eri vaiheisiin.

Vaihe	Ohjeita vaiheen toteuttamiseen
Mallinnusasetelman määrittäminen	Valitaan mallinnuksessa käytettävät muuttujat. Jos toteutetaan ennustava malli (ohjattu oppiminen), ennustamisen syötemuuttujina ei käytetä kiellettyjä syrjintäperusteita.
Datan esikäsittely	Varmista, että datan käsittelyssä on huomioitu anonymisointi ja pseudonymisointi asianmukaisesti. Toteuta datan jakaminen opetus-, testaus- ja validointijoukkoihin sopivassa jakosuhteessa. Jako opetusjoukkoihin tehdään mallin arkkitehtuurin ja kompleksisuuden valitsemiseksi ja hyvyyden realistisen arvioinnin toteuttamiseksi. Heikosti suunniteltu asetelma voi johtaa huonosti toimivaan tekoälymalliin. Toteutetaan muuttujien esikäsittely ja normalisointi. Normalisointi on tehtävä erikseen opetus-, testi- ja validointijoukoille. Datajoukkojen muutostenhallinnan toteuttamiseksi datajoukoilla tulee olla versionumerot ja versionhallinta.
Mallin opetus	Valitaan mallin rakenne ja arkkitehtuuri. Opetetaan malli tai joukko erilaisia malleja. Mallien muutostenhallinnan toteuttamiseksi malleilla tulee olla versionhallinta. Jos opetusdata sisältää henkilötietoa, selvitä differential privacy -tekniikoiden käyttämisen mahdolliset hyödyt. Tässä lähestymistavassa mallin opettamiseen lisätään satunnaisuutta, jonka tehtävä on taata henkilötietojen säilyvyyttä, ja estää henkilötietojen palauttamisen opetetusta mallista.
Mallin analysointi	Analysoidaan opetettujen mallien suorituskyky; ennustetarkkuus, sekä mahdollinen mallin reiluus (kts. 4.2 Reiluusmetriikat)

Asetelman ja mallin muokkaaminen analyysin perusteella	Asetelman muuttaminen, mallin arkkitehtuurin korjaaminen, datan esikäsittelyn muutos, vinoumien oikominen (kts. 4.3 Oikomismenetelmät)

4.1 Mallinnusvaiheen vinoumat

Opettamisen jälkeen mallien avulla voidaan muodostaa ennusteita. Turvallisen kehittämisen kannalta oleellista on tunnistaa ennusteiden mahdollisia vinoumia. Alla olevassa taulukossa on kuvattu yleisimpiä mallinnusvaiheen vinoumia, sekä ohjeita niiden käsittelyyn.

Taulukko 16: Mallinnusvaiheen vinoumia ja ohjeita niiden käsittelyyn.

Vinouma	Ohjeita
Syrjintäkriteereitä sisältävä malli	Opetettu malli tekee erilaisia ennusteita riippuen henkilöiden henkilökohtaisista ominaisuuksista. Henkilökohtaisia ominaisuuksia, kuten ikä, kieli, uskonto, sukupuoli, kansalaisuus, ei tule käyttää ennustavan mallin syötemuuttujina.
Välillinen syrjintä	Opetettu malli tekee erilaisia ennusteita riippuen henkilöiden henkilökohtaisista ominaisuuksista. Syötteenä ei tule käyttää myöskään sellaisia muuttujia, jotka välillisesti sisältävät edellä mainittua informaatiota.
Aggregointivinouma	Heterogeenisestä aineistosta muodostetaan keskiarvoistava malli, joka ei edustakaan aineiston kohderyhmiä. Esimerkiksi palkan ennustaminen työkokemuksen perusteella todennäköisesti vaatisi erilaiset mallit eri toimialoille. Yksi malli ei ehkä edusta ilmiötä minkään toimialan kohdalla.

4.2 Reiluusmetriikat

Turvallisen kehittämisen kannalta tärkeimmät työvaiheet ovat henkilötietojen hyödyntämiseen liittyvien ennustemallien tasapuolisuuden tai reiluuden varmistaminen. Reiluusmetriikat pyrkivät tunnistamaan, toteuttaako malli systemaattisesti erilaista kohtelua muodostaessaan ennusteita tai luokitteluita. Reiluusmetriikoita on kolmea kategoriaa:

- Yksilöreiluus: kohteleeko malli samalla tavalla henkilökohtaisilta ominaisuuksiltaan erilaisia, mutta muilta ominaisuuksiltaan samanlaisia yksilöitä
- Ryhmäreiluus: ryhmien välisten erojen tutkiminen tilastollisin testein
- Syrjintäperusteiden vaikutusta ennusteisiin arvioiviin menetelmiin

Reiluusmetriikoita on esitetty sekä regressio-, luokittelu- että klusterointimenetelmille. Turvallisen kehittämisen kannalta oleellista on tiedostaa, mitkä reiluusmetriikat soveltuvat erilaisiin sovelluksiin, ja tarpeen vaatiessa ratkaisun toteutuksessa voidaan toteuttaa myös tilanteeseen paremmin sopivia metriikoita (1).

Taulukko 17: Reiluusmetriikoita mallin hyvyyden mittaamiseen.

Metriikka	Kuvaus
Tilastollinen pariteetti	Positiivisen luokittelun todennäköisyys on yhtä suuri eri kohderyhmissä.
Ehdollinen tilastollinen pariteetti	Positiivisen luokittelun todennäköisyys on yhtä suuri kahdessa eri syrjintäperusteiden avulla määritetyssä ryhmässä, jos ne saavat samoja muuttujan arvoja. Laina myönnetään samalla todennäköisyydellä sekä naimissa olevalle että naimattomalle, jos heidän tulonsa ovat samat.
Virhetasojen yhtäläisyys	Väärän positiivisen ja väärän negatiivisen ennusteen todennäköisyys tulee olla sama esim. etnisestä ryhmästä riippumatta.
Yksilöreiluus	Kaikkien kahden yksilön parien ennusteiden ero riippuu ainoastaan hyväksyttävissä olevista eroista yksilöiden välillä. Ero etnisessä taustassa ei saa vaikuttaa ennusteissa esiintyviin eroihin.
Kalibraatio	Tavoitemuuttujan todellisten arvojen tulee olla yhtä suuria etnisestä ryhmästä riippumatta.

4.3 Oikomismenetelmät

Ennustavan mallin opettamisessa lopputuloksen hyvyyteen vaikuttavat olennaisesti käytettävissä oleva data-aineisto, aineiston käsittely ennen opettamista, käytetyn mallin arkkitehtuuri sekä tavoitefunktio. Lisäksi erilaiset mallit voivat tuottaa erilaisia arvoja reiluusmetriikoiden valossa tarkasteltuna (1).

Taulukko 18: Vinoumien oikomismenetelmiä.

Menetelmä	Kuvaus
Muuttujien uudelleenjärjestely	Mallin syöte- tai ennustemuuttujiin voidaan tehdä muutoksia; voidaan lisätä tai poistaa muuttujia. Myös muuttujien mahdollista esikäsittelyä voidaan muuttaa.
Syötedatan muokkaus	Datan jakosuhdetta opetus-, testaus- ja validointiaineistoihin voidaan muuttaa. Jos jostain reiluuden kannalta relevantista kohderyhmästä tai ilmiöstä ei ole riittävästi havaintoja, voidaan generoida synteettistä keinotekoista dataa tasapainottamaan kohderyhmien esiintyvyyttä.
Mallin rakenteen tai algoritmin muuttaminen	Tekoälymalleissa mallin rakennetta, eli vapaiden parametrien määrää ja muita vastaavia parametreja voidaan muuttaa, jotta saavutetaan parempi ennustetarkkuus ja reiluus.
Mallin ennusteiden jälkikäsittely	Joissain tilanteissa reiluuden lisääminen voidaan saavuttaa esimerkiksi ennusteiden kynnysarvoja muuttamalla.

4.4 Selitysmenetelmät

Tekoälyn hyödyntämisen yksi suurimmista riskitekijöistä liittyy tekoälymallien black box -ominaisuuteen, eli siihen, että mallin toiminnan ymmärtäminen ja selittäminen voi olla hyvin haastavaa. Tällaisissa olosuhteissa mallin mahdollisesti epäjohdonmukainen toiminta täysin uuden tyyppisissä tilanteissa voi olla arvaamatonta. Mallin toimintaa paremmin ymmärtämällä mahdolliset riskit ei-toivotuille ennusteille voitaisiin proaktiivisesti tunnistaa ja eliminoida. Tekoälyn selitettävyyden parantamiseksi on kehitetty joukko menetelmiä, joiden avulla tekoälyn toteuttaman ennustelogiikan ymmärtäminen voi mahdollistua. Alla olevassa taulukossa on kuvattu kaksi eri selittämisen käyttötapausta ja ohjeita selittämisen toteuttamiseksi.

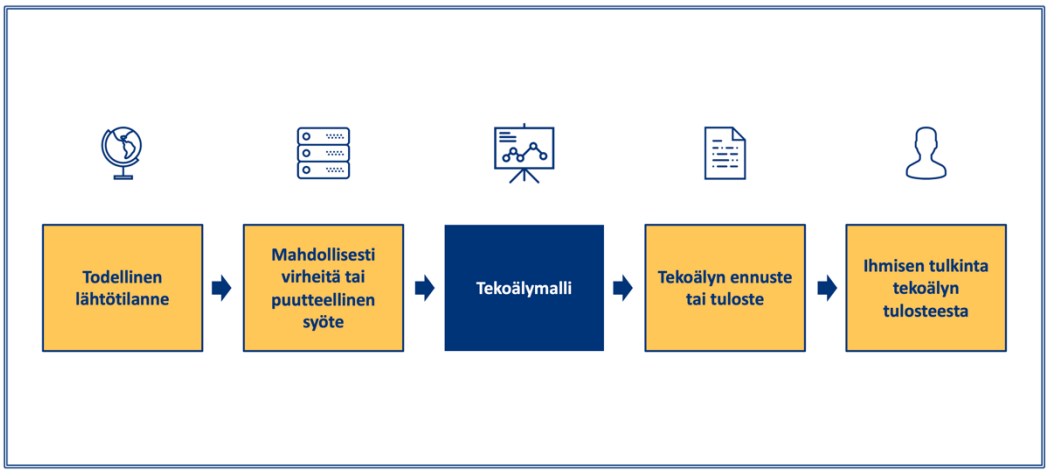
Taulukko 19: Tekoälymallien selitysmenetelmiä.

Käyttötapaus	Ohje
Tavoitteena on selittää tekoälymallin tekemä yksittäinen päätös numeerisen syötteen perusteella.	Selvitä yleisimpien selitysmenetelmien soveltuvuus käyttötapaukseen: SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-Agnostic Explanations). Huomaa, että esim. SHAP-menetelmää voi hyödyntää myös tekstimuotoiselle syötteelle.
Tavoitteena on ymmärtää koko aineiston laajuudessa mallin tekemiä päätöksiä.	Selvitä SHAP lähestymistavan sopivuus eri muuttujien tärkeyden selvittämiseen ennusteen muodostuksessa. Selvitä päätöspuihin perustuvien lähestymistapojen (Decision Trees) sopivuus selittämisen toteuttamiseksi.

4.5 Tekstipohjaisten mallien opettaminen

Kielipohjaiset käyttöliittymät yleistyvät voimakkaasti. Useimmiten ne perustuvat käyttäjän syöttämään tekstiin, mutta puheen tallentaminen ja muuttaminen tekstiksi myös suomeksi onnistuu hyvin. Kun tekoälyä kehitetään tekstiaineistojen avulla, törmätään uuden tyyppisiin haasteisiin.

Ilmeisimmät uhkakuvat tekstipohjaisten tekoälyratkaisujen käyttämiseen liittyvät tekstisyötteestä johdettuihin luokituksiin, suosituksiin tai päätöksiin, jotka voivat olla yksittäiselle käyttäjälle vääriä, vahingollisia tai syrjiviä. Kielen monimuotoisuuden takia toivotun ja tasapuolisen käsittelyn saavuttaminen ja takaaminen on haastavampaa kuin esim. numeerisen tiedon tapauksessa.



Kuva 5: Tekstin hyödyntämiseen perustuvan tekoälymallin toiminnan vaiheet.

Yllä olevassa kuvassa on esitetty tekoälypohjaisen tekstin käsittelyn vaiheet. Käsittely voi epäonnistua eri vaiheissa:

- Käyttäjän tuottama teksti tai puhe ei kuvaa käyttäjän lähtötilannetta oikein, tai sisällössä on esimerkiksi kirjoitusvirheitä
- Teksti tulkitaan tekoälyn toimesta väärin tai puutteellisesti
- Koneen tuottama vastaus tai ennuste on väärä tai arvaamaton

- Käyttäjä tulkitsee tai olettaa tuloksen virheellisesti

Alla olevassa taulukossa on kuvattu ohjeita eri vaiheiden toteuttamiseen.

Taulukko 20: Ohjeita tekstipohjaisten mallien toteuttamiseen.

Tilanne	Ohjeita
Käyttäjän syöttämä teksti ei välttämättä pidä paikkaansa.	Järjestelmän tulee kyetä toimimaan järkevästi myös väärällä ja arvaamattomalla syötteellä. Järjestelmän tulee pyrkiä validoimaan syötettä muiden tietojen avulla, mikäli mahdollista.
Järjestelmä ei ymmärrä syötettä oikein	Sanalistojen käyttäminen ja sanojen eksaktin esitysasun käyttäminen sääntöpohjaisesti voi johtaa arvaamattomaan lopputulokseen kirjoitusvirheistä ja sanojen sijamuodoista johtuen. Ngram -pohjaiset menetelmät sietävät kirjoitusvirheitä ja sijamuotoja eksaktia käsittelyä paremmin. Sanavektori- ja lausevektorimallit (word embedding) voivat auttaa järjestelmää tulkitsemaan oikein erilaisia samaa tarkoittavia sanoja tai lauseita. Malleja voi opettaa itse riittävän suurista aineistoista. Semanttisten mallien opettamisessa aineiston valinta on kriittistä. Internetin opetusaineisto voi sisältää yhteiskunnallista syrjintää. Usein asiayhteys on puutteellisesti ymmärretty. Kontekstin ymmärrys saattaa edellyttää teksti- / keskusteluhistorian sisällyttämistä käsittelyyn. Ongelmana voi olla, että käyttäjän kieltä tai murreta ei ymmärretä. Käytettävissä voi olla kielen kääntämisen palveluita tai kielimalleja, jotka ymmärtävät useita eri kieliä. Jos opetetaan chatbottia, esimerkkilauseita jokaiselle intentiolle on syötettävä riittävästi.
Järjestelmän tuottama luokittelu tai ennuste on syrjivä tai ongelmallinen	Järjestelmän yhtenä syötetietona on tekstiä, joka voi sisältää syrjintäkriteereitä tai syrjintäkriteerien kanssa korreloivaa informaatiota. Tällaisissa tilanteissa informaatio ei saa vaikuttaa lopputulokseen tai järjestelmän tuotokseen. Järjestelmää on testattava kattavasti. Sisällöllisesti sama käyttötapaus on testattava kielellisesti monipuolisesti. Järjestelmä tuottamien ennusteiden on kohdeltava eri kohderyhmistä johdettuja esimerkkejä reilusti ja tasapuolisesti. Järjestelmän täysin arvaamaton toiminta realistissa käyttötapauksissa tulee estää laajentamalla opetusaineistoa. Järjestelmää toteutettaessa on oltava laaja ymmärrys käyttäjäkunnasta ja heidän mahdollisista taustoistaan, ja eri tilanteisiin on kyettävä varautumaan.

Käyttäjä ymmärtää tai tulkitsee tuloksen tai tuotoksen virheellisesti	Käyttäjää on informoitava, että tuotos on tietojärjestelmän eikä ihmisen tuottama. On pyrittävä varmistamaan, että käyttäjä tiedostaa tuotosten olevan koneellisesti tuotettuja ja virheiden mahdollisuus on huomattava. Järjestelmä ei saa esittää tietoja tavalla, joka voi johtaa käyttäjää harhaan. Erityisesti laadukas, järjestelmän tuottama teksti voidaan helposti tulkita ihmisten tuottamaksi, mikä voi oleellisesti johtaa käyttäjää harjaan.

4.6 Valmismallien käyttäminen

Monissa sovelluksissa ei ole mahdollista tai tarkoituksenmukaista opettaa tekoälymallia itse opetusaineiston avulla. Esimerkiksi puheen tunnistuksessa tai kuvan tunnistuksessa voi olla mahdollista käyttää kolmannen osapuolen opettamia malleja. Myös erilaisten tekstipohjaisten valmismallien määrä kasvaa nopeasti. Valmiiksi opetetut mallit voivat tarjota useita hyötyjä. Esimerkiksi opettamiseen vaadittavan opetusaineiston määrä voi olla niin suuri, että sen hallinta organisaatiossa ei onnistu. Myös opettamiseen käytettävissä oleva laskentateho voi ylittää organisaation kyvykkyydet. Näissä tilanteissa valmismallit voivat olla hyvä vaihtoehto.

Valmiiksi opetetut mallit sisältävät myös vaaroja. Valmismallien tapauksessa organisaatiolla on alentunut kyky kontrolloida mallin toimintaa eri tilanteissa. Mallien toiminnan muokkaaminen voi myös olla vaikeaa tai mahdotonta. Nykyään monet valmismallit tarjoavat mahdollisuuden virittää mallia käyttäjän tarpeisiin. Alla esitetyssä taulukossa on kuvattu organisaation hallinnan kattavuus eri vaihtoehdoissa.

Taulukko 21: Organisaation erilaiset hallinnan tasot valmismallien hyödyntämisessä.

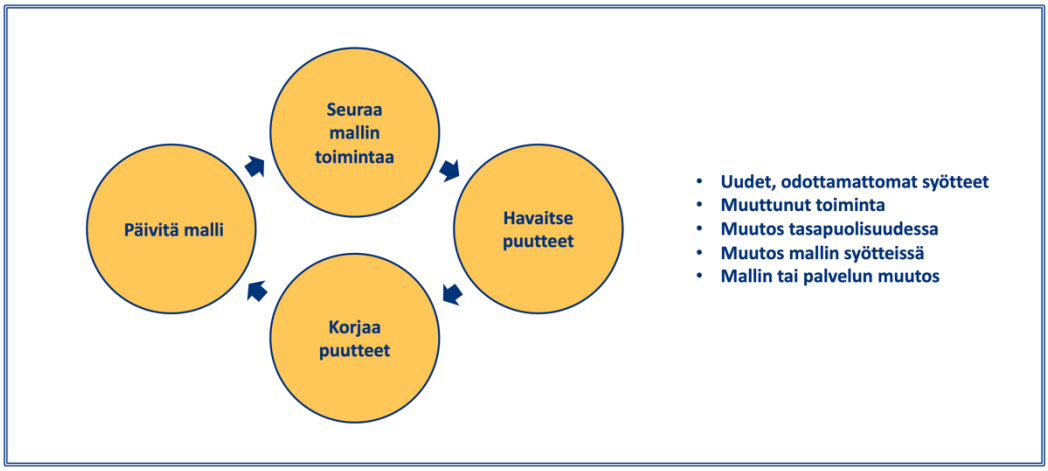
	End-to-end hallinta	Osittainen ratkaisun hallinta	Vähäinen ratkaisun hallinta
Sovelluksen hallinta	Organisaatio hallitsee tekoälyä hyödyntävää sovellusta	Organisaatio hallitsee tekoälyä hyödyntävää sovellusta	Organisaatio hallitsee tekoälyä hyödyntävää sovellusta
Mallin arkkitehtuurin hallinta	Organisaatio valitsee arkkitehtuurin	Organisaatiolla kopio mallista	Mallin arkkitehtuuri 3. osapuolen tiedossa
Mallin opetuksen ja datan hallinta	Organisaatio opettaa oman mallin omalla datalla	3. osapuoli opettaa mallin	3. osapuoli opettaa mallin

Alla kuvatussa taulukossa on esitetty ohjeita erilaisiin valmismallien yhteydessä esiintyviin haasteisiin.

Taulukko 22: Ohjeita valmismallien käyttämiseen.

Käyttötapaus	Ohjeita
Ratkaisu hyödyntää valmiiksi opetettua tekoälymallia.	Selvitä mallin käyttämiseen liittyvät ehdot ja huolehdi, että mallin hyödyntämiseen on suunnitelma ratkaisun koko elinkaaren ajaksi. Kiinnitä erityistä huomiota mallin toiminnan testaamiseen, ja erilaisten ryhmien reilun ja tasapuolisen kohtelun varmistamiseen.
Valmismalli ei toimi kaikilta osin täysin toivotusti	Selvitä, onko mallia mahdollista päivittää tai virittää opettamalla omalla aineistolla.

5 Käyttöönotto



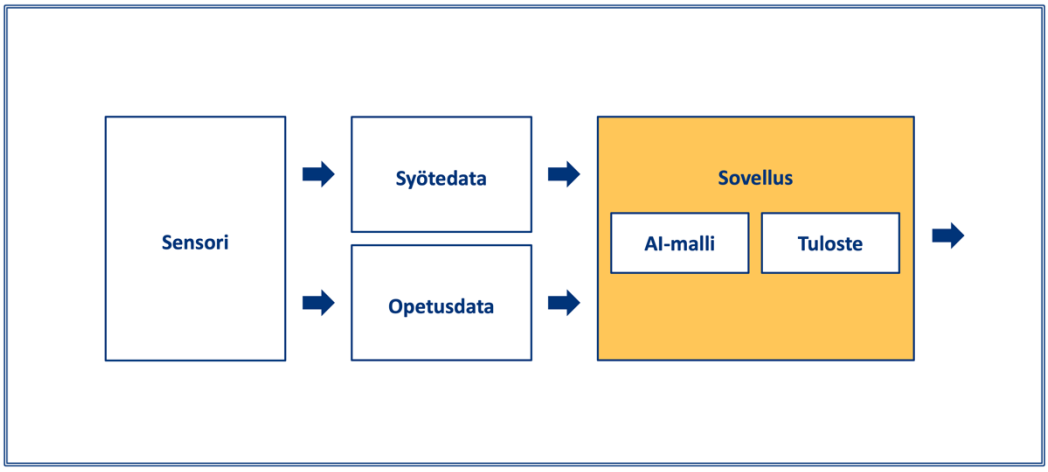
Kuva 6: Tekoälyn käyttöönoton vaiheet.

Tuotantoratkaisun toteuttamisessa pääpaino siirtyy jatkuvan mallin toiminnan ja sen muutosten seuraamiseen ja mallin jatkuvan päivittämisen kyvykkyyden kehittämiseen. Yllä esitetyssä kuvassa on kuvattu käyttöönoton vaiheet. Tuotantokäytön mahdollistamisen avaintekijöitä ovat mallin toiminnan automaattisen regressiotestaamisen järjestäminen, datan vinoumien ja mallin reiluuden jatkuva seuranta, sekä selitettävyydessä tapahtuvien muutosten jatkuva seuranta. Kehitysympäristön kannalta on huomioitava, että kehittämiseen voi osallistua useita henkilöitä, jolloin kumuloituvien muutosten kokonaisvaikutusten seuraaminen on ensisijaisen tärkeää. Ratkaisuun tehtävien muutosten tulee olla dokumentoituja ja jälkikäteen jäljitettäviä. Päivitetyn mallin julkaiseminen edellyttää metriikoiden huolellista tarkastelua sekä vastuuhenkilön selkeää päätöksentekoa ja sen dokumentointia.

Taulukko 22: Tekoälyn käyttöönottoon liittyvät ohjeet.

Ongelmatilanne	Ohjeita
Tekoälymallia sovelletaan täysin uusissa tilanteissa, joita ei esiintynyt mallinnusvaiheessa.	Laajenna opetusdataa kattamaan uudet ilmiöt. Laajenna testitapauksia kattamaan uudet ilmiöt, testitapaukset dokumentoivat minkälaista toimintaa uusissa tilanteissa odotetaan. Malli tulee päivittää kattamaan paremmin todellisuudessa esiintyvät tapaukset.
Mallin päivityksen jälkeen aiemmin hyvin toimineet tapaukset ovat muuttuneet mutta niitä ei havaittu ajoissa.	On kehitettävä mekanismi, jotta mallinkehityksen aikana havaitaan aiemmin toimineiden ominaisuuksien muuttuminen. Regressiotestauksen käyttötapausten määrää tulee laajentaa. Mallin uusia ominaisuuksia kehitettäessä vanhojen toiminnallisuuksien on edelleen toimittava suunnitellusti.
Mallin tasapuolisuus erilaisia ryhmiä kohtaan on muuttunut päivitetyssä mallissa, reiluusmetriikkojen arvot muuttuvat	Reiluusmetriikat tulee laskea ja niissä tapahtuva muutos tulee havaita. Malli on korjattava ennen päivitystä.
Mallin selitettävyydessä tai toiminnassa havaitaan muutos ei-toivottuun suuntaan.	Muutos selitettävyydessä eli mallin toiminnan kuvaamisessa on havaittava, ja ei-toivottu muutos on korjattava.
Mallin opetusdatassa tai mallin ennusteissa havaitaan uusia vinoumia.	Muutos vinoumissa havaittava, ja malli on korjattava ennen päivitystä.
Mallin syöte- tai ennustemuuttujat muuttuvat, ja syrjintäkriteereitä tulee mukaan syötteeseen.	Muutokset mallin muuttujissa ja rakenteessa tulee havaita ajoissa, syrjintäkriteerien tai niiden kanssa korreloivien muuttujien käyttö tulee havaita ja poistaa ennen päivitetyn mallin julkaisua
Mallin toiminta vaikuttaa käyttäjien toimintaan, yhteiskunnallista vinoumaa vahvistava toiminta havaitaan.	Selvitä, vahvistaako tekoälyn käyttäminen yhteiskunnallista tiedostettua vinoumaa. Selvitä, onko mallilla negatiivista kierrettä vahvistava vaikutus.
Mallin toimintaa, opetusdataa ja muuttujia muokataan ilman asianmukaista dokumentaatiota.	Mallin muutokset on dokumentoitava ja muutokset on kyettävä jäljittämään myöhemmin. Käytettävät mallit ja opetusaineistot on versioitava ja dokumentaation on sisällytettävä kuvaus toiminnasta eri versioissa.
Tekoälyratkaisu hyödyntää kolmannen osapuolen valmismalleja, ja niiden toiminta muuttuu.	Valmiskomponenttien ja niiden toteuttamien mallien toiminnan muutos tulee havaita. Automaattisen testaamisen tulee havaita muutos suhteessa aiempaan toimintaan.
Tuotannossa olevan mallin, ja sen käyttämiseksi rakennetun järjestelmän, tarvitsemista ohjelmistokirjastoista löytyy haavoittuvuuksia.	Toteutuksissa käytetään tyypillisesti kolmansien osapuolien kirjastoja. Ne vaativat ylläpitoa, koska haavoittuvuuksia löytyy, ja niitä korjataan jatkuvasti. Haavoittuvuuksien havaitsemiseksi ja korjaussuosituksien saamiseksi kannattaa ottaa käyttöön käytänteitä (ja mahdollisesti ohjelmistoja), joilla ohjelmointivirheet havaitaan mahdollisimman nopeasti.

5.1 Tekoälyjärjestelmiin kohdistetut hyökkäykset



Kuva 7: Tekoälyjärjestelmän komponentit.

Merkittävä osa tekoälyn hyödyntämisen riskeistä liittyy mallin kehittämiseen ja julkaisuun sekä opettamisessa käytettävän datan arkaluonteisuuteen. Tekoälyjärjestelmät ovat ohjelmistojärjestelmiä, joihin liittyy myös tietoturvallisuuteen liittyviä riskejä, kuten esimerkiksi mahdolliset ulkopuolisten suorittamat hyökkäykset. Yllä olevassa kuvassa on kuvattu tyypillisen tekoälyjärjestelmän komponentteja, joita kohtaan voidaan kohdistaa hyökkäyksiä. Alla olevassa taulukossa on kuvattu yleisimpiä hyökkäysten muotoja ja ohjeita hyökkäyksiltä suojautumiseen. Tekoälyjärjestelmiin kohdistuvien hyökkäysten kuvaaminen perustuu Traficomin julkaisuihin (9), (10).

Taulukko 23: Tekoälyjärjestelmiin kohdistuvat uhat ja niihin varautuminen.

Uhka	Ohjeita
Tekoälymallin vuotaminen ja varastaminen.	Varmista opetetun mallin tietoturvallisuus. Malliin pääsy on sallittu vain asianmukaisille henkilöille. Varmista, että mallin rekonstruointi syöte-vastaus pareista ei ole mahdollista, esim. rajoittamalla mallille tehtävien kyselyiden määrää. Mallin opettamisen hajautus; jos malli opetetaan osissa eri tietojen osalta, yksittäisen mallin vuotaminen ei paljasta koko ratkaisun toimintaa.
Opetusdatan vuotaminen	Varmista opetusdatan tietoturvallisuus. Varmista datan siirtämisessä käytettyjen tiedonsiirtokanavien turvallisuus. Differential privacy: dataan mahdollisesti lisätty kohina vaikeuttaa yksittäisten tietoalkioiden paljastumista.
Mallin takaisinmallinnus opetusdatasta tai esimerkeistä	Estä opetusdatan vuotaminen. Rajoita mallille tulevien kyselyiden määrää, jotta takaisinmallinnukseen vaadittavaa tietomäärää ei voida muodostaa.

Mallin väistäminen: syötteen manipulointi tekoälyn harhauttamiseksi	Vihamielinen opettaminen: opetusdataan lisätään muokattuja esimerkkejä, jolloin mallin sietokyky muokkauksille paranee. Syötteen ja ennusteiden tarkistaminen: poikkeavat syötteet tai mallin ennusteet voivat olla merkki manipuloidusta syötedatasta.
Opetusdatan myrkyttäminen: opetusdataa muokataan tahallisesti mallin toiminnan häiritsemiseksi	Regularisointi: rajoita mallin kompleksisuutta (esim. neuroverkon koko), eli käytä yksinkertaisia mallien arkkitehtuureja. Yksinkertaisemmat mallit eivät ole niin herkkiä ylisovittumaan syötedataan, mikä tekee niistä vikasietoisempia. Differential privacy: syötedataan lisätty kohina lisää mallin sietokykyä ja riippuvuutta yksittäisistä myrkytetyistä syötteistä.
Havaintodatan muuttaminen	Varmista, että järjestelmän syötettä ei ole mahdollista manipuloida.
Mallin käytön estäminen ja viivästäminen	Paikallinen päätöksenteko käyttäjän laitteessa: jos päätökset tehdään keskitetysti, luottamuksellisen tiedon siirtämisen tarve kasvaa, päätöksentekoon kuluva aika kasvaa. Aikakriittisissä sovelluksissa paikallinen päätöksenteko nopeuttaa päätöksen tuottamista. Laitteiden jatkuva käyttöasteen seuranta voi paljastaa palvelunestohyökkäyksen käynnistymisen.
Sensoreiden käytön estyminen	Varmista, että tekoälyjärjestelmän vaatiman syötteen keräämisen ja mittaamisen estäminen ei mahdollista koko järjestelmän toimintakyvyn kyseenalaistamiseen.

6 Tekoäly luovuutta ja tehokkuutta parantavana työkaluna

Tekoälyn hyödyntäminen ei liity yksinomaan tekoälyä hyödyntävien järjestelmien toteuttamiseen. Tekoälyä voidaan hyödyntää myös perinteisempien ohjelmistojärjestelmien toteuttamisessa sekä luovassa että asiantuntijaosaamista vaativissa tehtävissä. OpenAI:n julkaiseman ChatGPT -palvelun saama laaja huomio ja menestyminen erilaisissa tehtävissä luonnollisesti houkuttelee palvelun kokeilemiseen ja mahdolliseen hyödyntämiseen. Suurten kielimallien kehittyessä organisaatioiden olisikin aiheellista määritellä vastaavien kielimallien hyödyntämisen säännöt ja periaatteet. Erityisen tärkeää on tiedostaa suurten kielimallien heikkoudet ja niihin liittyvät riskit.

Alla olevassa taulukossa on kuvattu erilaisia kielimallien ja generoivan tekoälyn käyttötapauksia sekä ohjeita riskien välttämiseen.

Taulukko 24: Ohjeita tekoälyn hyödyntämiseen asiantuntijatyössä.

Käyttötapaus	Ohjeita
Ohjelmointikielen lähdekoodin generointi tekstisyötteen avulla	Huomioi, että annettu tekstisyöte lähetetään kolmannelle osapuolelle. Tekstisyöte saattaa sisältää luottamuksellista tietoa. Älä lähetä mitään aineistoa, josta esimerkiksi yritys on tunnistettavissa. Älä lähetä mitään henkilö- tai asiakastietoa tekstisyötteellä toimiviin tekoälypalveluihin. Älä lähetä mitään järjestelmä-, tietokanta- tai salasanatietoa tekoälypalveluihin. Ohjelmistokehittäjän tulee ottaa vastuu kehittämästään ohjelmakoodista. Huomioi, että tekoälyn generoima lähdekoodi voi sisältää virheitä.
Tiedonhaku, tekstin analysointi tai tekstin tiivistäminen kielimallien avulla	Huomioi, että tekoälypalveluille lähetetty tekstisyöte voi sisältää luottamuksellista tietoa. Älä lähetä mitään aineistoa, josta esimerkiksi yritys on tunnistettavissa. Älä lähetä mitään henkilötietoa tekstisyötteellä toimiviin tekoälypalveluihin. Älä lähetä mitään järjestelmä-, tietokanta- tai salasanatietoa tekoälypalveluihin. Tekoälyn generoimassa tekstissä voi olla hyvinkin merkittäviä asiavirheitä. Sisällön oikeellisuuden tarkistaminen on aina ihmisen ja asiantuntijan vastuulla. Huomioi, että generoitu teksti voi joissain tilanteissa vastata opetusaineiston tekstiä (lähes) sellaisenaan. Tekstin käyttämiseen voi liittyä tekijänoikeuskysymyksiä.
Kuvien generointi tekstisyötteen avulla	Huomioi, että generoidut kuvat voivat joissain tilanteissa vastata opetusjoukossa esiintyviä kuvia. Kuvien käyttämiseen voi liittyä tekijänoikeuskysymyksiä.
Kielikäännösten tekeminen	Huomioi, että annettu tekstisyöte lähetetään kolmannen osapuolen haltuun. Tekstisyöte saattaa sisältää luottamuksellista tietoa. Älä lähetä mitään henkilö- tai asiakastietoa tekstisyötteellä toimiviin tekoälypalveluihin.

7 Lähdeluettelo

1. **Ojanen, Atte;ym.** Algoritminen syrjintä ja yhdenvertaisuuden edistäminen : Arviointikehikko syrjimättömälle tekoälylle. [Online] 2022. http://urn.fi/URN:ISBN:978-952-383-404-0

2. **Digi- ja väestötietovirasto.** Turvallisen sovelluskehityksen käsikirja. [Online] 2023. https://wiki.dvv.fi/pages/viewpage.action?pageId=230470940.

3. **DigiFinland.** Virta-arkkitehtuuri. [Online] 2021. https://digifinland.fi/wp-content/uploads/2021/04/Virta-arkkitehtuuri_11.3.2021.pdf.

4. **Euroopan unioni.** EU:n yleinen tietosuoa-asetus. [Online] 2016. https://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CELEX:32016R0679&from=FI.

5. **Euroopan komissio.** Tekoälyasetus. [Online] 2021. https://eur-lex.europa.eu/legal-content/FI/TXT/HTML/?uri=CELEX:52021PC0206.

6. **Euroopan komission perustama asiantuntijaryhmä**. Luotettavaa tekoälyä koskevat eettiset ohjeet. [Online] 2019. https://api.hankeikkuna.fi/asiakirjat/ff3444f4-24c9-4ee8-8c9d-7bc581c0021a/796dac3f-4527-45c0-a7b8-d63024345ac8/JULKAISU_20200214084153.pdf.

7. **Oikeusministeriö.** Automaattiseen päätöksentekoon liittyvät yleislainsäädännön sääntelytarpeet - esiselvitys. [Online] 2020. https://api.hankeikkuna.fi/asiakirjat/ff3444f4-24c9-4ee8-8c9d-7bc581c0021a/796dac3f-4527-45c0-a7b8-d63024345ac8/JULKAISU_20200214084153.pdf.

8. **Eduskunta.** Laki hallintolain muuttamisesta, asian ratkaiseminen automaattisesti. [Online] 2023. https://www.finlex.fi/fi/laki/alkup/2023/20230487.

9. **Traficom.** Tekoälyn soveltamisen kyberturvallisuus ja riskienhallinta. [Online] 2021. https://www.traficom.fi/sites/default/files/media/publication/Teko%C3%A4lyn%20soveltamisen%20kyberturvallisuus%20ja%20riskienhallinta.pdf.

10. **Traficom.** Tekoälyn mahdollistamat kyberhyökkäykset. [Online] 2022. https://www.traficom.fi/sites/default/files/media/publication/TRAFICOM_Teko%C3%A4lyn_mahdollistamat_kyberhy%C3%B6kk%C3%A4ykset%202022-12-12_web.pdf.