VTT

Julkishallinto tukemassa eettisen tekoäly-yhteiskunnan rakentamista

Katsaus tekoälyn ohjauskeinoihin ja politiikkatoimenpiteisiin

Anton Sigfrids | Mika Nieminen | Jaana Leikas | Antero Karvonen | Pietari Pikkuaho

VTT TECHNOLOGY 421

VTT TECHNOLOGY 421

Julkishallinto tukemassa eettisen tekoäly-yhteiskunnan rakentamista

Katsaus tekoälyn ohjauskeinoihin ja politiikkatoimenpiteisiin

Anton Sigfrids, Mika Nieminen, Jaana Leikas,

Antero Karvonen & Pietari Pikkuaho

VTT

ISBN 978-951-38-8784-1

VTT Technology 421

ISSN-L 2242-1211
ISSN 2242-122X (Verkkojulkaisu)
DOI: 10.32040/2242-122X.2023.T421

JULKAISIJA – PUBLISHER

VTT
PL 1000
02044 VTT
Puh. 020 722 111
https://www.vtt.fi

VTT
P.O. Box 1000
FI-02044 VTT, Finland
Tel. +358 20 722 111
https://www.vttresearch.com

Alkusanat

Tekoälyn edistysaskeleet herättävät kysymyksiä tekoälysovellusten yhteiskunnallisista vaikutuksista. Monet valtiot, yritykset ja kansainväliset organisaatiot ovat kehittäneet ohjauskeinoja, joilla on pyritty lieventämään tekoälyyn liittyviä riskejä ja maksimoimaan sen hyödyntämisen tuomat edut.

Ohjaus- ja koordinaatiokeinoja kehitettäessä keskeisiksi kysymyksiksi nousevat miten haasteita ja riskejä olisi hallittava, millaisten arvojen perusteella toimitaan, millaisia tavoitteita asetetaan ja millaisten institutionaalisten mekanismien ja periaatteiden avulla tavoitteet voidaan saavuttaa. Tekoälyn ohjaus- ja koordinaatiokeinot kehittyvät nopeasti kansallisilla ja kansainvälisillä foorumeilla ja myös tutkimuskirjallisuutta julkaistaan kiihtyvään tahtiin. Tekoälyn hallinta etsii muotoaan.

Tässä raportissa koostamme ja tarkastelemme tutkimuskirjallisuudessa esiintyviä ehdotuksia julkisen hallinnon ohjaus- ja koordinaatiomekanismien parantamiseksi. Kiinnitämme erityistä huomiota sellaisiin periaatteisiin ja keinoihin, joiden avulla julkishallinto voi ohjata tekoälyn kehittäjiä ja käyttäjiä omaksumaan eettisiä ja vastuullisia käytäntöjä. Tarkastelumme osoittaa, että julkishallinnon tulisi omaksua osallistavan päätöksenteon muotoja pyrkimyksissään kokonaisvaltaisempaan ja koordinoituun tekoälyn ohjaukseen, valvontaan, sekä käyttötilanteittain räätälöityihin eettisiin toimintamalleihin.

Ehdotamme ratkaisuksi ideaalityyppistä OSKI-mallia (Osallistava, Soveltuva, Kokonaisvaltainen ja Institutionalisoitu), jossa ehdotettujen kehittämisratkaisujen keskeiset ulottuvuudet yhdistyvät tekoälyn käytön kokonaisvaltaiseksi ohjausmalliksi. Tämän tekstin aiempi versio on julkaistu englanninkielisenä artikkelina Sigfrids ym. (2022).

Raportti on toteutettu strategisen tutkimuksen neuvoston rahoittamassa ETAIROS-tutkimushankkeessa, jonka keskeinen tavoite on vahvistaa julkisen ja yksityisen sektorin yhteistä ymmärrystä tekoälykehityksen eettisistä käytännöistä ja pelisäännöistä. ETAIROS kehittää eettisesti kestäviä tekoälyn suunnittelumenetelmiä ja tuottaa näkemystä tekoälyä koskevan koordinaation ja ohjauksen keinoista yhdessä sidosryhmien kanssa.

Sisällysluettelo

Keskeiset käsitteet

Ennakointi	Innovaatiotoiminnan tai teknologisen kehittämistyön vaikutusten ennakointi talouden, yhteiskunnan/ihmisten ja ympäristön kannalta.
Teknologian etiikka	Tarkastelee teknologian suunnitteluun, kehittämiseen, hyödyntämiseen ja käyttöön liittyviä eettisiä kysymyksiä.
Tekoälyn etiikka	Tarkastelee älykkäiden teknologioiden (kuten itseohjautuvien ajoneuvojen, kasvojentunnistuksen, robottien tai konekääntäjien) kehitystä ohjaavia arvoja.
Tekoälykehitys	Tässä raportissa: tekoälyn käytön kehittyminen ja siitä johtuva sosiotekninen muutos yhteiskunnassa. Oleellisia kysymyksiä ovat mm. tekoälypohjaisten ratkaisujen yhteiskunnallinen haluttavuus ja hyväksyttävyys sekä ympäristön kestävyys.
Tekoälyn kehittäminen	Tässä raportissa: tekoälytekniikan tutkimus, kehitys ja innovaatiotoiminta. Oleellisia kysymyksiä ovat mm. tekoälyn luotettavuus, vastuuvelvollisuus, läpinäkyvyys, selitettävyys ja tietosuoja.
Ohjaus	Tässä raportissa: erilaiset toimintaa jäsentävät suunnittelun, päätöksenteon, ohjauksen ja organisoinnin muodot. Katso kansainvälisessä kirjallisuudessa käytetty käsite ”governance”.
Tekoälyn ohjaus	Tässä raportissa: Tekoälykehitykseen vastaavat ja tekoälyn kehittämistä jäsentävät suunnittelun, päätöksenteon, ohjauksen ja organisoinnin muodot.
Vastuullinen tutkimus- ja innovaatiotoiminta	Responsible Research and Innovation (RRI); Tutkimus- ja kehitystoimintaa ja tuloksia kehitetään ja arvioidaan vastaamaan yhteiskunnallisia arvoja, tarpeita ja odotuksia yhteistyössä tutkijoiden ja sidosryhmien kanssa. RRI-periaatteita ovat ennakointi, mukaan ottaminen, itsereflektio ja valmius muuttua.

Tiivistelmä

Tekoälystä on niin Suomessa kuin kansainvälisesti tulossa yleiskäyttöteknologia, jonka nähdään tuottavan hyötyjä monella eri alalla. Tekoälysovelluksiin liittyy kuitenkin tahattomia ja ristiriitaisia vaikutuksia, jotka voivat olla haitallisia yksilöille tai yhteiskunnalle. Nykyiset julkiset ohjauskeinot eivät ehkäise tekoälyn käytön nostattamia sosiaalisia riskejä ja uhkia riittävästi, minkä vuoksi niitä tulee kehittää.

Organisaatioilla ja julkishallinnolla tulisi olla kyvykkyyttä ennakoida ja tunnistaa tekoälyn riskejä kokonaisvaltaisesti ja vastata ketterästi syntyviin ongelmiin. Onnistuessaan julkinen ohjaus luotsaa tekoälyn kehittäjiä ja käyttäjiä siten, että tekoäly tuottaa yhteistä hyvää niin yksilöiden, yhteisöjen kuin koko yhteiskunnan kannalta. Tämä tapahtuu koordinoimalla politiikkatoimia nykyistä selvemmin sekä tukemalla vastuullista toimintaa ja kyvykkyyttä hahmottaa tekoälyn eettisiä vaikutuksia systeemisesti.

Esittelemme tässä raportissa systemaattisen kirjallisuuskatsauksen perusteella (Sigfrids ym. 2022) laatimamme tekoälyn julkisen ohjauksen tukemiseksi tarkoitetun OSKI-mallin (Osallistava, Soveltuva, Kokonaisvaltainen ja Institutionalisoitu). OSKI-malli kuvaa ideaalitilaa, ja tarjoaa julkisohjauksen suunnittelulle ja kehittämiselle suuntaviivat, joita kohti pyrkiä.

- **Osallistava**. Tekoälyn haasteiden hahmottaminen ja ohjauksen kehittäminen edellyttää yhteistyötä ja yhteiskehittämistä eri sidosryhmien, kuten tutkijoiden, yritysten, hallinnon virkamiesten, kansalaisyhteiskunnan ja kolmannen sektorin kesken. Yhteisten ratkaisujen löytämiseksi tulee arvioida toimenpiteiden vaikutuksia eri sidosryhmille sekä luoda edellytykset käydä keskustelua mahdollisista eturistiriidoista ja eri arvojen toteutumisesta. Eri sidosryhmien näkemysten ja arvojen huomioiminen vuorovaikutteisessa prosessissa on ketterien ja sosiaalisesti hyväksyttävien ohjauskeinojen ja eettisten periaatteiden suunnittelun ja toimeenpanon lähtökohta. Monialaista yhteistyötä voidaan vahvistaa kehittämällä vastuullisen tekoälyn ekosysteemejä sekä perustamalla toimintaa tukeva organisaatio tai elin.
- **Soveltuva**. Yhteiskunnallisesti tavoiteltavien, kestävien ja eettisesti hyväksyttävien ratkaisujen löytymiseksi voidaan soveltaa vastuullisen tutkimus- ja innovaatiotoiminnan (eng. Responsible Research and

Innovation, RRI) käytäntöjä, sekä jo olemassa olevia tekoälyn etiikkaa ja vastuullisuutta tukevia suunnittelumenetelmiä ja työkaluja. Lisäksi pitää tehostaa tekoälyn vaikutusten ennakointia ja arviointia. Eri sidosryhmien kyvykkyyttä soveltaa näitä työkaluja ja arvioida tekoälyn eettisiä ulottuvuuksia tulee tältä osin tukea esimerkiksi kehittämällä alan koulutusta. Tarvitaan myös konkreettisia käytännön ohjeita ja keinoja eettisen periaatteiden ja ihmisoikeuksien huomioimiseksi tekoälyä koskevassa päätöksenteossa.

- **Kokonaisvaltainen**. Parhaiden ratkaisujen löytämiseksi ja uusien kehittämiseksi tarvitaan muutokseen vaikuttavien tekijöiden kokonaisvaltaista ymmärtämistä sekä tekoälyn systeemisen luonteen ja pitkän aikavälin vaikutukset huomioiva näkökulma. Kokonaisvaltaisen näkökulman omaksuva julkinen hallinto huomioi esimerkiksi tekoälyn sovellusalueet, eettiset ja lainsäädännölliset kysymykset sekä keskeisten toimijoiden näkökulmat, arvot ja intressit. Ratkaisut tulee sovittaa niin pitkäaikaisiin strategioihin kuin paikallisiin tarpeisiin, ja niiden kehittämisessä tulisi huomioida tekoälyn kehittäjien, loppukäyttäjien sekä päätöksentekijöiden näkökulmat.
- **Institutionalisoitu**. Osallistavan päätöksenteon valmistelun järjestäminen tulisi olla valtakunnallisen elimen vastuulla, joka koordinoi tekoälypolitiikkaa ja kehittää ohjauskeinoja, kouluttaa sidosryhmiä, kerää tarpeellisen tiedon ja tuo eri sidosryhmät keskinäiseen vuoropuheluun. Lisäksi tarvitaan vakiintunutta viranomaistahoa valvomaan ja varmistamaan lainsäädännön noudattamista ja tukemaan eettisten normien toteutumista.

Toimenpide-ehdotukset

Tekoälyn julkista ohjausta tukevan OSKI-mallin ideaalit ja siihen liittyvät toimenpide-ehdotukset ovat toisiaan tukevia ja täydentäviä. Alla esitettyjen ehdotusten tarkoituksena on rohkaista keskustelua erilaisista toimenpiteistä, jotka tukevat vastuullisempaa tekoälyn hyödyntämistä Suomessa.

Luodaan tutkimustietoon perustuva kokonaisvaltainen tekoälyn hallintamalli:

- Kokonaisvaltaisuuden edistämiseksi tutkijoiden tulisi yhdessä virkahenkilöiden sekä muiden sidosryhmien kanssa kehittää kokonaisvaltaisten ratkaisujen kehittämistä tukeva hallintamalli. Mallin tarkoituksena on auttaa päätöksentekijöitä koordinoimaan ja sovittamaan tekoälyn paikalliset sovellusalueet, niiden eettiset ja lainsäädännölliset kysymykset sekä keskeisten toimijoiden näkökulmat, arvot ja intressit yhteen pitkäaikaisten vaikutustenarviointien ja kestävyysstrategioiden kanssa.

Perustetaan osallistavaa päätöksentekoa tukeva ja tekoälypolitiikkaa koordinoiva valtakunnallinen elin (Koskimies ym. 2021):

- Osallisuuden edistämiseksi tulisi perustaa monialaista yhteistyötä, eri sidosryhmien osallisuutta ja tekoälypoliittista keskustelua tukeva organisaatio tai elin. Organisaation tehtäviin kuuluisi luoda yhteinen keskustelufoorumi, jossa käydään moniäänistä sidosryhmäkeskustelua esimerkiksi teollisuuden ja kansalaisyhteiskunnan edustajien, valtiollisten toimijoiden, standardointielimien ja muutoin aliedustettujen ihmisryhmien välillä. Sidosryhmäkeskustelun tarkoituksena on tuottaa ehdotuksia ja suosituksia sosiaalisesti, ekologisesti ja taloudellisesti kestävän tekoälypolitiikan suunnittelemiseksi. Kyseinen organisaatio tukisi julkista hallintoa sopeutumaan tekoälykehityksen nopeaan tahtiin ja siitä nouseviin erilliskysymyksiin ja voimistaisi vastuullisuusperiaatteiden hyödyntämistä politiikkatoimenpiteiden suunnittelussa. Organisaatio voisi myös tukea julkisen keskustelun synnyttämistä tekoälyä ja digitalisaatiota koskevista strategioista ja tekoälyä ohjaavista periaatteista Suomessa.

- Tutkimuslaitosten tulisi yhdessä viranomaisten, demokratiatutkijoiden ja tuotekehittäjien kanssa pilotoida uusia digitaalisia kansalaisosallisuuden sekä sidosryhmäosallisuuden muotoja osana nopeasti kehittyvien teknologian yhteiskunnallista ohjausta, hyvää hallintoa ja tietoon perustuvaa päätöksentekoa. Pilotteja tulisi tutkia ja parhaita oppeja tulisi soveltaa käytännössä.
- Tekoälyn odotettavissa olevia hyötyjä voidaan kasvattaa ja riskejä minimoida koordinoimalla eri politiikkatoimia nykyistä selvemmin. Tekoälyn soveltamiseen, ohjaukseen ja sääntelyn on muodostettava johdonmukainen linja, johon on voitava luottaa investointeja suunnitellessa ja T&K-panoksia käytettäessä. Tätä tukemaan tarvitaan valtakunnallinen elin, jonka tehtävänä tulisi olla eettisesti kestävän ja vastuullisen tekoälypolitiikan muotoileminen, toimenpidesuositusten antaminen sekä niiden koordinoiminen ja seuranta. Elimen tehtävänä olisi esimerkiksi koordinoida toimintaa hallitusohjelmien, kansallisen tekoälyohjelmien, rahoitusinstrumenttien sekä näitä toimeenpanevien tahojen välillä, sekä tukea vastuullista tekoälyn käyttöä mahdollistavia paikallisia ekosysteemejä, joissa on käytössä esimerkiksi vankat auditointi- ja sertifiointimekanismit. Lisäksi viranomaisen on varmistettava, että sillä on riittävä kyvykkyys säännellä, valvoa ja seurata tekoälyjärjestelmiä nopeasti muuttuvassa toimintaympäristössä, ja että on olemassa selkeät vastuuketjut tekoälyn aiheuttamien ongelmien edessä.

Luodaan hyvät edellytykset vastuulliselle toiminnalle panostamalla koulutukseen, informaatio-ohjaukseen ja kannustimien luomiseen:

- Eri sidosryhmien kyvykkyyttä toimia eettisesti ja vastuullisesti voidaan tukea panostamalla alan koulutukseen ja kannustamalla toimijoita omaksumaan olemassa olevia vastuullisen tekoälyn ohjeita ja toimintamalleja (kuten AI HLEG 2019, UNESCO 2021, OECD 2023). Kirjallisuus tarjoaa laajan kirjon ohjeita esimerkiksi kansalaisosallisuuden tukemiseen, arviointityökalujen suunnittelemiseen ja taloudellisten kannustimien luomiseen.

1 Johdanto

Viimeisten vuosien aikana tehdyt harppaukset koneoppimisessa, saatavilla olevan datan määrän räjähdysmäinen kasvu, erilaisten datapankkien ja tietokantojen yleistyminen sekä prosessoreiden tehon kasvu ovat edistäneet merkittävästi tekoälyn hyödyntämismahdollisuuksia. Kehityksen seurauksena tekoälyksi luonnehdittua teknologiaa sovelletaan yhä enenevässä määrin monilla yhteiskunnan aloilla ja odotukset tekoälyn jatkuvan kehittämisen ja käyttöönoton suhteen ovat korkealla. Samaan aikaan tekoälyn nopean kehityksen kanssa on kansainväliseen keskusteluun noussut koulutusdataan ja algoritmeihin perustuvien järjestelmien haavoittuvuus suhteessa länsimaissa tärkeänä pidettyihin ja yleisesti hyväksyttyihin periaatteisiin. Tekoälyn laajempi käyttöönotto on synnyttänyt erilaisia riskejä ja ongelmia. Tutkijat ovat esittäneet, että tekoälyn hyödyntäminen tehokkuuden, optimoinnin ja voiton maksimoimiseen saattaa lisätä yhteiskunnallista eriarvoisuutta ja yksityisten toimijoiden yhteiskunnallista vaikutusvaltaa, kaventaa ihmisten autonomiaa, voimistaa ympäristötuhoa ja romuttaa ihmisten luottamusta digitaalisiin palveluihin (Stahl 2021; Crawford 2021, Zuboff 2019). Tekoälyn hyödyntämiseen liittyvien ongelmien välttämiseksi ja mahdollisuuksien valjastamiseksi monet valtiot, kansainväliset järjestöt ja yritykset sekä tutkijat ovat kehittäneet ja ehdottaneet uudenlaisia keinoja tekoälyn ohjaamiseksi[1], esimerkiksi eettisten periaatteiden, teollisuusstandardien ja lainsäädännön avulla. Ohjauskeinojen on tarkoitus luotsata tekoälyn käyttöönottoa toivottuun kehityssuuntaan samalla minimoiden tekoälyn ongelmia ja riskejä.

Tekoälyn ohjauskeinojen tutkimus perustuu ajatukseen, että olemassa olevia ohjauskeinoja olisi sovellettava paremmin ja kehitettävä edelleen, jotta julkiset hallintoelimet voivat vastata tehokkaasti tekoälyyn liittyviin haasteisiin samalla varmistaen, että tekoäly hyödyttää yhteiskuntaa laajasti ja tuottaa yhteistä hyvää (Stahl 2021; Ireni-Saban & Sherman 2021; de Almeida ym. 2021; Ulnicane ym. 2021; Larsson 2020; Nieminen ym. 2019; Floridi ym. 2018; Wirtz ym. 2020; Yeung ym. 2019; Taeihagh 2021; Truby 2020). Nykyisiin tekoälyn ohjauskeinoihin kuuluvat erilaiset soveltuvat säädökset, alan standardit, eettiset säännöt ja ohjeet, toimintastrategiat sekä sidosryhmien välistä koordinointia ja yhteistyötä tukevat

[1] ”Tekoälyn ohjauksella” viittaamme kansainvälisessä kirjallisuudessa käytettyyn käsitteeseen ”governance”, jolla tarkoitetaan tässä yhteydessä laajasti erilaisia toimintaa jäsentäviä suunnittelun, päätöksenteon, ohjauksen ja organisoinnin muotoja.

menettelyt. Vaikka lainsäädäntö ja tekniset standardit muodostavat tärkeän osan ohjauskeinoista, ne eivät yksinään riitä ohjaamaan tekoälykehitystä yhteiskunnallisesti tarkoituksenmukaiseen ja hyödylliseen suuntaan, minkä vuoksi tarvitaan jatkuvaa eettistä pohdintaa ja arviointia (Floridi 2018). Laki tarjoaa karkeat puitteet kehitykselle. Merkittävä osa eettisesti merkityksellisestä toiminnasta tapahtuu näissä laajoissa raameissa: laki ei pysty säätelemään toiminnan kaikkia piirteitä. Tekniset standardit tarjoavat myös yhden näkökulman säätelyyn, mutta sen vaikutukset laajemmassa eettisessä mielessä ovat aina osa laajempaa sosioteknistä kokonaisuutta, jossa teknologiaa sovelletaan ja hyödynnetään erilaisiin tarkoituksiin vaihtelevissa inhimillisissä puitteissa. Standardit tarjoavat lain tapaan puitteita, joissa teknologiaa hyödynnetään. Ohjauksen tulisi puolestaan palvella tämän kokonaisuuden kehittämistä siten että kaikki erilaiset ohjauskeinot ja niiden suhde ohjauksen tavoitteisiin on jatkuvan tarkastelun ja kehittämisen kohteena.

Tekoälyn eettinen ohjaus pyrkii ohjaamaan tekoälyn kehitystä ja käyttöä kohti sosiaalisesti hyväksyttäviä ja toivottavia sekä yhteiskunnallisesti hyödyllisiä päämääriä (Stahl 2021; Ireni-Saban & Sherman 2021; Mazzi & Floridi 2023; Taddeo & Floridi 2018; Winfield & Jirotka 2018). Tekoälyn valjastaminen näiden päämäärien työvälineeksi edellyttää tekoälyn eettisten periaatteiden sulauttamista ohjaus- ja koordinaatiomekanismeihin ja edelleen konkreettiseen toimintaan tekoälyn kehittäjien ja käyttäjien keskuudessa. Eettisten periaatteiden heikko täytäntöönpano yritystoiminnassa sekä julkishallinnon politiikkatoimenpiteissä on kuitenkin saanut osakseen kritiikkiä (Stix 2021; Larsson 2020; Hagendorff 2020; Morley ym. 2020; Mittelstadt 2019). Ongelmana ei ole eettisten periaatteiden tai niihin liittyvien toimenpidesuositusten puute, vaan niiden muotoilu käytännössä sovellettaviksi ja vaikuttaviksi keinoiksi niin yritysten itsesääntelyn keinoina kuin politiikkatoimenpiteissä (Stahl ym. 2021; Yeung ym. 2019; Stix 2021).

Tekoälyn ohjauksen ja koordinaation kehittämistä hankaloittavat yhteiskunnan kompleksisuus sekä tekoälyinnovaatioiden uutuus, laaja-alaisuus ja nopea kehitysvauhti. Eräänä ratkaisuna uusien teknologioiden kehitystyön ja käytön ohjaukseen on esitetty niin kutsuttua ennakoivaa ja mukautuvaa ohjausta (esim. Ulnicane ym. 2021; Taeihagh 2021; Taeihagh ym. 2021), joka perinteisten hierarkkisten lähestymistapojen sijaan tukee verkostomaista, sidosryhmiä osallistavaa, ja asiantuntevaan tietoon perustuvaa innovaatioiden ohjausta (Guston 2014; Kuhlmann 2019; Taeihagh, Ramesh & Howlett 2021). Julkinen hallinto voi toiminnallaan edistää ennakoivan ohjauksen periaatteiden toteutumista tukemalla erilaisten sidosryhmien välistä vuorovaikutusta, konsensuksen löytymistä, yhteistä tiedontuotantoa, ja ohjauskeinojen jatkuvaa yhdessä kehittämistä (Lähteenmäki-Smith 2020, Borras & Edler 2020). Niin sanotut joustavat eli ennakoivat ja mukautuvat ohjauskeinot edellyttävät siirtymistä hierarkkisista ohjauskeinoista kohti koordinointi-, ennakointi- ja osallistumisprosesseja korostavia menetelmiä (Taeihagh ym. 2021; Kuhlmann ym. 2019). Tämä on keino parantaa kollektiivista päätöksentekoa ja lisätä julkishallinnon ohjauksen joustavuutta ja sopeutumiskykyä uusien teknologioiden yhteiskunnallisiin vaikutuksiin.

Tekoälyn ohjausta voi siis tarkastella kahdesta osin päällekkäisestä näkökulmasta. Ensimmäinen keskittyy päätöksenteon menettelyiden kehittämiseen (Kuhlmann ym. 2019; Taeihagh ym. 2021) ja toinen eettisten periaatteiden vaikuttavuuteen (AI HLEG 2019; Cath ym. 2017; Jobin ym. 2019; Floridi ym. 2018). Vastuullisen tutkimuksen ja innovoinnin (eng. Responsible Research and Innovation RRI) näkökulmasta niin esitykset joustavista teknologian menettelyistä ja eettisistä periaatteista näyttävät lähekkäisiltä (Ireni-Saban & Sherman 2021; Winfield & Jirotka 2018; Lehoux ym. 2020). Molemmissa lähestymistavoissa hyödynnetään vastuullisen tutkimuksen ja innovoinnin periaatteita (Ireni-Saban ja Sherman 2021; Winfield & Jirotka 2018; Lehoux ym. 2020). RRI-lähestymistavan mukaan sidosryhmien osallistuminen, vuoropuhelu ja erilaisten näkökulmien huomioon ottaminen innovaatioiden varhaisessa vaiheessa ovat olennaisen tärkeitä sen varmistamiseksi, että erilaiset yhteiskunnalliset arvot ja intressit tulevat huomioiduiksi teknologiaa koskevassa päätöksenteossa.

Tekoälyn ohjaus- ja koordinaatio (eng. governance of artificial intelligence) on laaja ja nopeasti kehittyvä kokonaisuus. Tekoälyn julkiseen ohjaukseen liittyvää tutkimuskirjallisuutta ja kansallisia strategioita julkaistaan kiihtyvällä vauhdilla (esim. Zuiderwijk ym. 2021; Wamba ym. 2021) ja kansainvälisistä näkökulmaeroista huolimatta kirjallisuus tarjoaa useita mahdollisia vastauksia vastuullisen ja eettisen tekoälykn edistämiselle. Tutkijat ovat tarkastelleet esimerkiksi ohjausta kohti yhteistä hyvää (Stahl 2021; Floridi ym. 2020; Wamba ym. 2021; Tomašev ym. 2020) ja kestävää kehitystä (Truby 2020; Vinuesa ym. 2020). Tutkijat ovat myös esittäneet tekoälyn ohjaukselle ja koordinaatiolle universaalia, ihmisarvoihin perustuvaa arvopohjaa (Yeung ym. 2019; Smuha 2020; Donahoe & Metzger 2019). Ehdotukset voidaan jakaa kolmeen ryhmään: politiikkatason ja lainsäädännön kehittämisen ehdotukset (Djeffal ym. 2022; Stix 2021; Stahl 2021), julkisen ja yksityisen sektorin organisaatioille kohdistetut ehdotukset (Mäntymäki ym. 2022; Zuiderwijk ym. 2021; Shneiderman 2020), sekä käytännönläheiset työkalut ja ohjeet, joiden tarkoituksena on tukea esimerkiksi tekoälyn kehittäjiä ja päättäjiä eettisemmän ja vastuullisemman tekoälyn toteuttamisessa (Morley ym. 2020; Zicari ym. 2022; UNESCO 2021). Tässä raportissa tarkastelemme politiikkatason ja käytännön ehdotuksia julkisen hallinnon ohjaus- ja koordinaatiomekanismien parantamiseksi ja eettisten toimintamallien edistämiseksi.

Vaikka tekoälyn ohjausta ja koordinaatiota koskevassa kirjallisuudessa tarjotaankin erilaisia ehdotuksia eettisen ja vastuullisen tekoälyn käytön edistämiseksi, kirjallisuus on hajanaista ja on epäselvää, miten julkishallinto voisi hyödyntää esitettyjä ehdotuksia. Kysymme tässä raportissa, miten julkishallinto voisi tukea tekoälyyn perustuvien palveluiden kehittäjiä ja soveltajia omaksumaan eettisiä ja vastuullisia periaatteita toiminnassaan. Esittelemme tässä raportissa Sigfrids ym. (2022) systemaattisen kirjallisuuskatsauksen tuloksia tekoälyn julkisen ohjauksen konkreettisista keinoista sekä uudistamisen tarpeista.

Tietojemme mukaan vain muutama tutkimus on koonnut tutkimuskirjallisuudessa esiintyviä ehdotuksia julkishallintoa koskevien toimenpiteiden osalta (Wirtz ym. 2020; Stahl 2021). Aiheesta on julkaistu vain yksi systemaattinen katsaus (de Almeida ym. 2021). de Almeidan tutkimuksessa katsastetaan 21 tutkimusartikkelia

ja poliittista asiakirjaa, jonka perusteella, mutta tutkimuksen analyysiä perustelematta, johdetaan tekoälyn julkisen ohjauksen hallintamalli[2]. Mallissa tutkijat ehdottavat julkisen hallinnon instituutioille uusia tehtäviä, mutta jättävät käsittelemättä hallinnon uudistamisen tarpeet, mukaan lukien sidosryhmien osallisuuden kehittämisen ja eettisen ja vastuullisen innovaatiotoiminnan toteutumisen edellytysten parantamisen. Käsillä olevassa raportissa pyrimme sen sijaan kiinnittämään huomiota ”joustavaan” julkiseen ohjaukseen, RRI lähestymistavan toimeenpanemiseen sekä laajaan sidosryhmäosallisuuteen politiikkavalmisteluissa ja päätöksenteossa, sillä nämä ovat eettisen tekoälyn ohjauksen kehittämiselle keskeisiä (Winfield & Jirotka 2018; Kuhlmann ym. 2019; Lehoux ym. 2020; Ireni-Saban & Sherman 2021; Taeihagh 2021; Ulnicane ym. 2021).

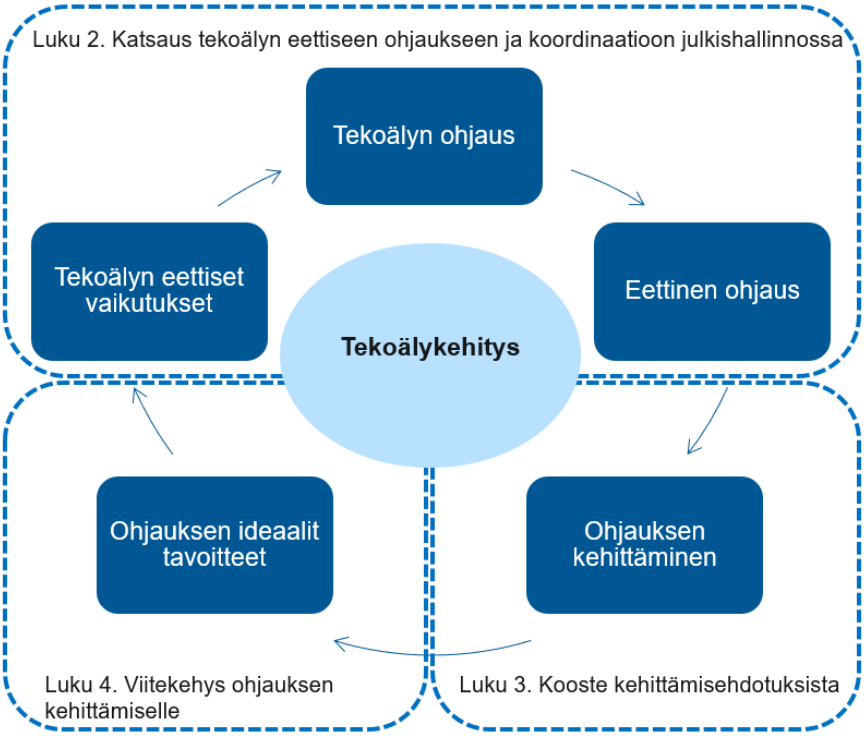
Tässä raportissa esitellään Sigfrids ja kumppaneiden (2022)[3] julkaiseman kirjallisuusselvityksen tuloksia. Selvityksessä analysoitiin 21 tutkimusartikkelin ehdotukset tekoälyn julkisen ohjauksen kehittämisestä (menetelmäkuvaus liitteessä 1) ja tulokset tiivistettiin neljään eettisen ohjauksen ideaalityyppiin. Näiden näkökulmien koostaminen auttaa laajentamaan keinovalikoimaa, jolla julkisen hallinnon on mahdollista edistää etiikan ja ihmisoikeuksien toteuttamista tekoälyä koskevassa poliittisessa päätöksenteossa sekä organisaatioiden käytännöissä (Ireni-Saban & Sherman 2021; Winfield & Jirotka 2018; Lehoux ym. 2020; Ulnicane ym. 2021; Taeihagh 2021; Kuhlmann ym. 2019). Raportissa esitettyjen kehitysehdotusten ja analysoidun kirjallisuuden perusteella voimme tiivistäen todeta, että tekoälyn julkisen eettisen ohjauksen tulisi hyödyntää etiikkaan, ihmisoikeuksiin, hyvään hallintoon sekä vastuulliseen tutkimukseen ja innovointiin perustuvia normatiivisia periaatteita politiikanteossa.

Kuvio 1 auttaa hahmottamaan raportin sisältöä. Luvussa 3 teemme yleiskatsauksen tekoälyn eettiseen ohjaukseen. Taustalla on ajatus, että tekoälysovellusten yleistymisen myötä niiden kehittämiseen ja käyttöön liittyy tahattomia ja ristiriitaisia vaikutuksia, jotka voivat olla haitallisia yksilöille ja yhteiskunnalle. Tekoälyteknologian käyttöönoton myötä syntyy täten uudenlaisia eettisiä kysymyksiä ja vastuualueita, joihin julkisen johtamisen on vastattava ja sopeuduttava. Tekoälykehityksen haltuunotto ohjauskeinojen avulla on kuitenkin haasteellista nopean kehitysvauhdin, vaikean ennakoitavuuden ja muutoksen laaja-alaisuuden vuoksi. Tekoälyn eettinen ohjaus tavoittelee yhteistä hyvää tukeutumalla eettisiin, ihmisoikeudellisiin, hyvän hallinnon sekä vastuullisen tutkimus- ja innovoinnin normatiivisiin periaatteisiin. Ongelmana on, että eettisen ohjauksen toimeenpano uupuu. Tämän vuoksi esittelemme neljännessä luvussa tutkimuskirjallisuuteen perustuvan koosteen tekoälyn ohjauskeinojen kehittämisehdotuksista. Raportin viimeisessä luvussa kokoamme tulokset yhteen kokonaisvaltaiseen malliin, jonka esitämme ideaalityyppisenä viitekehyksenä

[2] Uutta kirjallisuutta aiheesta julkaistaan jatkuvasti. Tämän raportin analyysissä ovat mukana huhtikuuhun 2021 mennessä julkaistut tutkimukset.

[3] Tämän raportin sisältö perustuu suurelta osin Sigfrids ym. 2022 julkaisemaan tieteelliseen artikkeliin, jota on päivitetty ja täydennelty.

julkishallinnon ohjaus- ja koordinaatiokeinojen kehittämiselle. Kirjallisuuskatsauksen rajaus ja menetelmät on kuvattu lyhyesti liitteessä 1.



Kuva 1. Raportin pääaiheet

2 Tekoälyn julkinen ohjaus ja koordinaatio

2.1 Mitä tekoälyn ohjaus on?

Englanninkielinen governance eli ohjaus- ja koordinaatio on moniulotteinen käsite, jolle on olemassa erilaisia määritelmiä (Frederickson 2007). Suomeksi puhutaan hallinta-ajattelusta tai ohjauksesta ja koordinaatiosta, joka yleisellä tasolla viittaa päätöksentekoon ja päätösten toimeenpanoon liittyvään toimintaan. Laajasti katsottuna käsite viittaa eri toimijoiden väliseen järjestäytyneeseen interaktioon eli formaaliin ja epäformaaliin sääntelyyn tai ohjaukseen, joka johdattaa toimintaa tai käyttäytymistä kohti asetettuja tavoitteita (Asaduzzaman & Virtanen 2016). Julkishallinnon kontekstissa hallinnan kehittämisen käsitteeseen liittyy tyypillisesti ajatus siirtymisestä pois jäykästä ja hierarkkisesta päätöksenteon rakenteesta kohti kansalaisia ja yhteiskunnan toimijoita osallistavaa ja joustavaa toimintaa. Puhutaan paradigman muutoksesta, jossa julkishallinto etsii uusia keinoja mukautua ja vastata yhteiskunnan lisääntyneeseen kompleksisuuteen ja sen aiheuttamiin "viheliäisiin ongelmiin" (esim. Torfing ym. 2020). Muutosta voidaan kuvata hallinnon siirtymisenä hierarkkisesta, sääntelyyn keskittyvästä hallinnosta kohti verkostoja ja osallistumista sekä vuorovaikutteista ja demokraattisempaa hallintoa. Uudenlainen ohjaus ja koordinaatio eivät korvaa sitä edeltäneitä hallinnan mekanismeja, vaan se on syntynyt näiden rinnalle. Näin hallintamekanismit yhdistävät perinteisiä ohjaus- ja sääntelymekanismeja uudenlaisen verkostomaisen toiminnan kanssa, johon liittyy esimerkiksi ilmiölähtöisyys, älykäs erikoistuminen sekä deliberaatio ja dialogi. Lähestymistavassa painottuu tulevaisuussuuntautuneisuus sekä erilaisten sidosryhmien osallistaminen ja verkostomainen koordinaatio, jonka mahdollistajana julkinen hallinto toimii. (Lähteenmäki-Smith ym. 2021).

Teknologian ohjauksessa (eng. technology governance) on yleisesti kyse erilaisten ohjauskeinojen, sääntelyn ja koordinaatiomekanismien soveltamisesta teknologian käyttöön ja kehitykseen. Floridi (2018, 3) määrittelee digitalisaation ohjauksen "politiikkatoimenpiteitä, menettelyitä ja standardeja toimeenpaneviksi käytännöiksi, joiden tavoitteena on tukea informaatioympäristön (eng. infosphere) asianmukaista kehittämistä, käyttöä ja hallintaa”. Tällaista ohjausta tukee hyvä koordinointi sekä digietiikka ja sääntely. Tekoälyn ohjauksesta ja koordinaatiosta (eng. governance of AI) ei ole olemassa yhtä selkeää määritelmää. Termin käyttö vaihtelee niin tutkimusjulkaisuissa (Zuiderwijk ym. 2021) kuin poliittisissa asiakirjoissa (Ulnicane ym. 2021, 78). Tyypillisesti käsitteellä viitataan tekoälyn

yhteiskuntaa muuttavan potentiaalin valjastamiseen ja riskien minimoimiseen erilaisten ohjauskeinojen avulla. Tällaisia mekanismeja ovat esimerkiksi sääntely, eettiset periaatteet, informaatio- ja resurssiohjaus, alan standardit sekä erilaiset tekniset ratkaisut ja ohjelmistoarkkitehtuurit (Stahl 2021; Morley ym. 2020; Clarke 2019).

Gahnberg (2021) on esittänyt, että tekoälyn ymmärtäminen sen toimijuuden kautta auttaa käsitteellistämään tekoälyn ohjauksen kohdetta. Tekoälyä voidaan tästä näkökulmasta ymmärtää keinotekoisena toimijana (eng. artificial agent), jolla on kyky ”hahmottaa ympäristöään antureiden avulla ja vaikuttaa siihen ohjaimilla” (eng. actuator) (Russell & Norvig 2009, 34), ja jonka voi määritellä tiettyjen perusominaisuuksien kautta. Tällainen toimijuus ilmentyy joko tietokoneohjelmistoissa tai fyysistesti roboteissa ja autonomisissa laitteissa. Perusominaisuuksia ovat (Gahnberg 2021, Russell & Norvig 2009): suoritemittaus (mitä tuloksia toimija tavoittelee), ympäristö (minkä ympäristön kanssa toimija on vuorovaikutuksessa), toimintatapa (eng. actuator, miten toimija vaikuttaa ympäristöönsä) ja havaitseminen (millä tavalla toimija saa tietoa ympäristöstään) (Gahnberg 2021, 196). Esimerkiksi autonomisen ajoneuvon suoritemittaus voi liittyä matkan turvallisuuteen, nopeuteen, lainmukaisuuteen tai mukavuuteen. Ajoneuvon ympäristöjä voivat olla tiet, muu liikenne, jalankulkijat ja matkustajat. Toimilaitteita voivat olla ohjauspyörä, kaasupoljin, jarru, äänitorvi ja näyttö. Havaitsemisen antureina puolestaan voivat toimia kamerat, valotutka, nopeusmittari, GPS ja moottorin sensorit (Gahnberg 2021, Taulukko 1).

Tekoälyn toimijuutta peilaaviin perusominaisuuksiin perustuen Gahnberg (2021, 195) päätyy määrittelemään tekoälyn ohjauksen ”intersubjektiivisesti tunnustettuina sääntöinä, jotka määrittävät, rajaavat ja muovaavat odotuksia keinotekoisen toimijan perusominaisuuksista”. Toisin sanoen, ohjauksen kohteena ovat keinotekoisen toimijan perusominaisuuksiin liittyvät kontekstisidonnaiset ilmentymät niin digitaalisissa kuin fyysisissä ympäristöissä. Suoritemittaukseen liittyvä hallinta ohjaa keinotekoisen toimijan tavoitteidenasettelua. Tähän liittyviä sääntöjä voivat olla esimerkiksi kansainväliset pyrkimykset kieltää ns. tappajarobotit, joiden tavoitteena on ihmisten haavoittaminen. Myös GDPR -asetukseen sisältyvä algoritmisen päätöksenteon selitettävyyden ja läpinäkyvyyden vaatimus ohjaa tavoitteenasettelua suuntaamalla toimijat tavoittelemaan suoritteita, joihin sisältyy selitettävyys onnistumisen mittarina. Mittareiden määrittämistä ja yhtenäistämistä tukee erilaisten teollisuusstandardien kehitys.

Keinotekoisen toimijan ympäristöä koskeva ohjaus voi tarkoittaa sellaisen alueen määrittelemistä, jossa autonomisella ajoneuvolla on sallittua ajaa. Alue voi olla myös digitaalinen, esimerkiksi online-pokerisivustot ovat kieltäneet erilaisten bottien hyödyntämisen sivustoillaan. Autonomisten bottien toimintatapaa ympäristössä voidaan puolestaan säädellä esimerkiksi kieltämällä automaattisia vastauksia, roskaposteja tai valeinformaatiota tuottavat botit digitaaliselta alustalta, kuten Twitter on tehnyt. Ajoneuvon toimintatapaa voidaan puolestaan säädellä esimerkiksi määrittämällä, millaisia toimintoja tietyillä alueilla voidaan automatisoida. Näitä voivat olla esimerkiksi ajoneuvon hallinnan osittainen automaattinen avustus. Yleiseen tietosuoja-asetukseen liittyvä yksityisyyden suoja

on hyvä esimerkki sääntelystä, joka kieltää keinotekoista toimijaa keräämästä tietynlaista dataa havaitsemisteknologian avulla, ellei siihen ole erillistä lupaa. (Gahnberg 2021, 201–205)

Kun tekoälyn ohjaus ja hallinnan kohde määritellään yllä kuvatusti keinotekoisten toimijoiden yleistettävien perusominaisuuksien perusteella, voidaan määrittelyn pohjalta todeta kenen tai minkä hallinnon vastuualueeseen kunkin perusominaisuuden paikallinen ilmentymä kuuluu, kuka siihen voi vaikuttaa ja minkälaisia intressejä ja toimijoita siihen liittyy. Gahnbergin ehdottama tekoälyn ohjauksen kehys keskittyy sääntelyn ja muiden ohjauskeinojen soveltamiseen tekoälytoimijoihin, mutta ei kuvaile menettelyjä, joiden kautta parhaita ohjaus- ja sääntelykeinoja voidaan kehittää, eikä sitä, miten tätä prosessia tukevaa vuorovaikutteista ja osallistavaa toimintaa voidaan järjestää. Jää myös epäselväksi, millä tavalla algoritmien ja koneoppimisen erilaisiin eettisiin ongelmiin voidaan tällaisen hallintakehikon kautta kiinnittää huomiota.

Tekoälyn ohjauksella viittaamme tässä raportissa tekoälyn kehittämistä ja käyttöä luotsaaviin ohjauskeinoihin, etiikkaan ja sääntelyyn. Ohjauskeinot viittaavat menettelyihin, koordinaatioon ja politiikkatoimenpiteisiin, joita viranomaiset ja poliittiset päättäjät hyödyntävät päätöksenteon muodostamiseen sekä yksityisten ja julkisten tekoälyn käyttäjien ja sidosryhmien ohjaamiseksi. Etiikka viittaa arvioihin toiminnan eettisyydestä ja eettisten periaatteiden toimeenpanoon. Sääntely puolestaan viittaa tapoihin, joilla lainsäädännön toimeenpano ja noudattaminen ovat osa ohjauskäytäntöjä. (ks. Floridi 2018).

> ***Tekoälyn ohjaus sisältää erilaisia toimintatapoja, menettelyitä ja sääntöjä, joiden tarkoituksena on ylläpitää ja tukea yhteisten arvojen muotoilemiseksi tarvittavia yhteistyön muotoja, sekä yhteistä päätöksentekoa tekoälyn kehittämisen tavoitteista ja niiden toimeenpanosta (ks. Dafoe 2018).***

Kuten johdannossa esitimme, tekoälyn eettinen ohjaus pyrkii ohjaamaan tekoälyn kehitystä ja käyttöä kohti sosiaalisesti hyviä ja yhteiskunnallisesti hyödyllisiä päämääriä (Stahl 2021; Ireni-Saban & Sherman 2021; Taddeo & Floridi 2018; Winfield & Jirotka 2018). Eettisen ohjauksen tavoitteet voivat olla osittain päällekkäisiä muiden tavoitteiden, kuten tehokkuuden lisäämisen, optimoinnin tai sosiaaliseen kontrollin kanssa (Stahl ym. 2021).

2.2 Ohjauksen haasteet ja keinovalikoimaa

Tekoälyn ohjaus toivottuun kehityssuuntaan sekä ohjaus- ja koordinaatio-toimenpiteiden kehittäminen on osoittautunut haasteelliseksi (katso kuva 3). Tämä johtuu siitä, että tekoälyn kehitys on nopeaa, vaikeasti ennakoitavissa, tapahtuu usealla yhteiskunnan osa-alueella samanaikaisesti ja tuottaa täysin uudenlaisia riskejä ja odottamattomia vaikutuksia. Taeihagh (2021) ja kumppaneiden mukaan uuden teknologian sääntelyä hankaloittaa tyypillisesti informaation asymmetriat, toimenpidepoliittinen epävarmuus, rakenteellinen valtadynamiikka sekä politiikkatoimenpiteiden virheet. **Informaation asymmetria** viittaa tilanteeseen,

jossa tieto uuden teknologian vaikutuksista sekä eri toimijoiden tavoitteista ja mahdollisuuksista jakautuu epätasaisesti ja voi johtaa ennalta arvaamattomiin ja ei-tarkoitettuihin vaikutuksiin. Eri toimijoilla on eri kompetenssit arvioida teknologian todellisia käyttötarkoituksia, ja tietoa omaavat tahot ovat etuasemassa muihin nähden. **Toimenpidepoliittinen epävarmuus** viittaa tilanteeseen, jossa hallinnolla ei ole täyttä ymmärrystä sääntelyä vaativasta ongelmasta, eikä varmuutta tarvittavasta ratkaisusta. **Rakenteellinen valtadynamiikka** viittaa tilanteeseen, jossa teknologian hyödyt jakaantuvat epätasaisesti väestöryhmien kesken, mahdollisesti voimistaen joidenkin sidosryhmien vaikutusvaltaa päätöksentekoon. **Virheelliset politiikkatoimenpiteet** ovat teknologian ohjauksessa tyypillisiä ja kiteytyvät tunnettuun Collingridge-ongelmaan. Sääntelevä viranomaisen on valittava, millä tavalla teknologian kehitystä ohjataan, vaikka saatavilla ei ole riittävästi tietoa teknologian laaja-alaisen käytön vaikutuksista. Viranomaisen on tasapainoteltava kansalaisia riskeiltä suojaavan ennakoivan sääntelyn ja vähemmän säännellyn ja siten mahdollisesti innovointia tukevan lähestymistavan välillä.

Kuva 2. Tekoälyn julkisen ohjauksen haasteita ja keinovalikoimaa.

Ohjauskeinot voidaan karkeasti ottaen jakaa kahteen eri luokkaan: pehmeään ja kovaan sääntelyyn (eng. soft vs hard law) (Gutierrez ym. 2021; Wallach & Marchant, 2018). Kovat ohjauskeinot ovat oikeudellisesti sitovia säädöksiä, jotka määrittävät mitä on sallittua tehdä ja mitä ei. Ne viittaavat lakeihin ja muihin sanktioituihin sääntöihin, joita lainsäätäjä kehittää ja toimeenpanee virallisten lainsäädäntöprosessien kautta. Pehmeät keinot puolestaan viittaavat erilaisiin eettisiin periaatteisiin, suosituksiin ja käytännesääntöihin sekä erilaisiin teknisiin järjestelmä- tai infrastruktuurirakenteisiin ja niihin liittyviin standardeihin ja protokolliin, joita käytetään usein organisaatioiden ja toimialojen itsesääntelyn muotoina (ks. Clarke 2019). Uusien teknologioiden ohjauksessa teknologian kehittäjät ja hyödyntäjät käyttävät tyypillisesti ensisijaisesti pehmeän sääntelyn välineitä organisaatioiden ja teollisuudenalojen itsesääntelyyn, ja lainsäädäntöä kehitetään tarvittaessa jälkikäteen (Taeihagh ym. 2021). **Katsomme pehmeän ohjauksen käsittävän myös laajemman joukon keinoja, kuten koordinoinnin**

sekä resurssi- ja informaatio-ohjauksen, joilla ohjataan teknologioiden kehittämistä ja käyttöönottoa. Jako pehmeän ja kovan sääntelyn sekä itsesääntelyn ja virallisen hierarkkisen sääntelyn välillä kuvastaa jatkuvaa jännitettä, joka syntyy toisaalta julkisen hallinnon tarpeesta säännellä prosesseja riskien välttämiseksi ja yhteiskunnallisten tavoitteiden edistämiseksi, ja toisaalta tarpeesta taata sääntelyn kohteena olevien tahojen autonomia.

Kansainvälisesti tekoälyä ohjaavat useat kovan ja pehmeän sääntelyn muodot. Kansainvälisen standardisoimisjärjestön ISO:n ja sähkö- ja elektroniikkainsinöörien instituutin IEEE:n kehittämät kansainväliset standardit ohjaavat tekoälyn kehittämistä teknisellä tasolla ja tekoälyn kehittäjien keskuudessa (Cihon 2019). Sekä EU (Euroopan komissio 2021a) että OECD (2019) ovat antaneet politiikkasuosituksia tekoälyn turvallisen ja hyödyllisen kehityksen avittamiseksi. Muita periaatteita ja suosituksia ovat muun muassa UNESCOn tekoälyn etiikkaa koskeva suositus Recommendation on the Ethics of Artificial Intelligence, jonka UNESCOn yleiskonferenssi hyväksyi istunnossaan 24. marraskuuta 2021, sekä ”Asilomar AI Principles”, IEEE:n ”Ethically Aligned Design”, ”Charlevoix Common Vision for the Future of Artificial Intelligence” (Charlevoixin yhteinen visio tekoälyn tulevaisuudesta), ”DeepMind Ethics & Society Principles” (DeepMindin eettiset ja yhteiskunnalliset periaatteet), ”Google AI Principles” (Googlen tekoälyperiaatteet) ja ”The Information Technology Industry AI Policy Principles” (ks. Future of Life Institute 2021).

Vaikka EU:ssa ja Yhdysvalloissa on tehty erilaisia ehdotuksia tekoälyä koskevaksi erityissääntelyksi (esim. nykyinen EU:n tekoälylakiehdotus, Euroopan komissio (EC), 2021b, ks. tekoälyä käsittelevä korkean tason asiantuntijaryhmä, EC 2022), tekoälyä koskevat erityissäännökset ovat vasta valmisteilla (The AI Act 2022). Näin ollen tekoälyä koskeva nykyinen lainsäädäntö koostuu erilaisista tekoälyä koskevista yleisluontoisemmista asetuksista. Näitä ovat muun muassa ihmisoikeudet EU:n perusoikeuskirjassa ja yksityisyyden suoja yleisessä tietosuoja-asetuksessa (GDPR) sekä tuotevastuudirektiivi, syrjinnän vastaiset direktiivit ja kuluttajansuoja. EU:ssa on vastikään hyväksytty kaksi uutta lainsäädäntöehdotusta koskien digipalveluita (Digital Services Act, DSA) ja digitaalimarkkinoita (Digital Markets Act, DMA). Näillä päivitetään digitaalipalveluja koskevia EU:n laajuisia sääntöjä ja tavoitteena on sekä suojella digitaalisten palvelujen käyttäjien perusoikeuksia että edistää innovointia, kasvua ja kilpailukykyä.

Vaikka sääntely ja standardit voivat olla tehokkaita tekoälyn ohjauskeinoja, ne eivät välttämättä riitä ohjaamaan tekoälyä yhteiskunnallisesti tarkoituksenmukaiseen tai yhteisen hyvän suuntaan, minkä vuoksi tarvitaan eettistä näkökulmaa ja kykyä soveltaa eettistä päätöksentekoa eri yhteyksissä (Floridi 2018; Delacroix & Wagner 2021). Tästä johtuen eettiset periaatteet ja ohjeistukset ja säännöt ovat nousseet keskeiseksi tekoälyn pehmeiksi ohjauskeinoiksi (Jobin ym. 2019; Floridi ym. 2018; Stix 2021). Tekoälyn ohjauskirjallisuudessa ei löydy tarkkaa määritelmää tekoälyn eettisetä ohjauksesta. Yleisesti ottaen sillä viitataan ohjaus- ja koordinaatiokeinoihin, joiden tavoitteena on minimoida tekoälyn riskejä ja tukea teknologian käyttöä yhteisen hyvän sekä sosiaalisen ja taloudellisen kestävyyden hyväksi (Stahl 2021; Ireni-Saban & Sherman 2021). Hyvän hallintotavan periaatteet,

kuten hallinnon tehokkuus, läpinäkyvyys, osallisuus, vastaanottavaisuus ja palautteenantokyky, ja lainmukaisuus, liittyvät läheisesti myös eettiseen ohjaukseen (Winfield ja Jirotka, 2018, 2). Tekoälyn eettinen ohjaus perustuu osin myös vastuulliseen tutkimus- ja innovaatioitoimintaan (RRI; Ireni-Saban & Sherman 2021; Winfield & Jirotka 2018), jonka tavoitteena on monenlaisia sidosryhmiä osallistamalla varmistaa, että innovaatiotoiminta on yleisen edun mukaista.

Tekoälyn kehittämisen ja käytön tueksi laadittuja eettisiä periaatteita ja muita teollisuuden itsesääntelyn pehmeitä ohjauskeinoja on kritisoitu hampaattomina dokumentteina, joiden ei ole käytännössä mahdollista juurtua organisaatioiden tai innovaatiopolitiikan ohjaus- ja koordinaatiomekanismien osaksi (Larsson 2020; Hagendorff 2020; Morley ym. 2020; Mittelstadt 2019; Stix 2021; Clarke 2019; Delacroix & Wagner 2019; Yeung ym. 2019). Tekoälyn eettisen ohjauksen haasteena ei välttämättä ole yhteisten arvojen tai eettisten periaatteiden puute. Kyse on pikemminkin niiden kokoamisesta käytännössä sovellettavissa ja operationalisoitavissa olevaksi kokonaisuudeksi niin tekoälyä koskevassa poliittisessa päätöksenteossa kuin tekoälysovellusten ja -palvelujen kehittämisessä (Stahl ym. 2021; Donahoe ym. 2019).

2.3 Tekoälyn sosiotekniset vaikutukset

Tekoälyn ohjauskeinoja kehittävien ja soveltavien toimijoiden kannalta haasteena on, että tekoälyn erilaisista määritelmistä on ohjaskeinoja koskevassa kirjallisuudessa käyty vain vähän keskustelua (Gahnberg 2021; Taeihagh 2021; Clarke 2019b). Tekoälylle ei ole yhtä vakiintunutta määritelmää. Tyypillisesti termillä viitataan algoritmeihin, laskennallisiin tekniikoihin tai tietokonejärjestelmiin, joita voi hyödyntää tukemaan tai korvaamaan ihmistyötä vaativia tehtäviä esimerkiksi päätöksenteossa, asiantuntijajärjestelmissä, kuvien tunnistuksessa, kielen kääntämisessä ja sen tuottamisessa. Niin sanottua keinotekoista älykkyyttä, eli koneoppimista tai neuroverkkoja hyödyntäviä järjestelmiä tuotetaan pitkäjänteisen opettamisen, laajojen datamäärien ja ohjelmistoarkkitehtuurien avulla, erilaisia sääntö- ja palkintomekanismeja hyödyntäen. Tavoitteena on laskennallinen järjestelmä, joka ympäristöään havainnoiden ja siihen joustavasti sopeutuen pyrkii maksimoimaan annettujen tavoitteiden toteutumista (Christian 2020). Tekoälyä on sen etiikkaan ja hallintaan liittyvässä kirjallisuudessa tulkittu toimijuuden kautta: interaktiivisena, autonomisena ja osin itseoppivana toimijana, joka voi suoriutua ihmisälykkyyttä ja ihmistyötä tyypillisesti vaativista tehtävistä (Floridi 2020; Taeihagh 2021).

Tällaiselle teknologialle on monia eri käyttömahdollisuuksia, minkä vuoksi sitä voi myös kuvailla yleiskäyttöteknologiana, joka mahdollistaa uudenlaisten tuotteiden ja toimintatapojen kehittämisen. Tällaisia tuotteita ovat esimerkiksi automaatioon perustuvat järjestelmät, lohkoketjuteknologiat sekä asioiden internetiin ja virtuaalitodellisuuteen perustuvat tuotteet. Tekoälyä käytetään esimerkiksi terveydenhuollossa diagnosoinnin ja terveyden ylläpitämisen apuvälineenä, rikosten ennaltaehkäisemisessä, osakemarkkinoilla ja tietokonepeleissä. Tekoälyä voidaan hyödyntää julkisen sektorin palveluiden parantamisen ja saavutettavuuden sekä julkishallinnon lainsäädännön ja päätöksenteon tukena (Sharma 2020; Zuiderwijk ym. 2021).

Tekoälyn kohdalla tehdään usein erottelu “kapean” ja “yleisen” tekoälyn välillä. Kapea tekoäly (ANI, artificial narrow intelligence) tarkoittaa tekoälyä, joka on kehitetty ratkaisemaan vain tiettyjä rajattuja tehtäviä. Näitä voivat olla esimerkiksi kasvojentunnistus, auton ohjaaminen tai tuotteen suosittelu asiakkaalle. Kapean tekoälyn näennäinen “älykkyys” on kuitenkin täysin rajattu tehtävään, johon se on kehitetty (Naudé & Dimitri 2020). Kapean tekoälyn toiminta perustuu algoritmeihin, jotka voivat hyödyntää suuria määriä dataa “oppiakseen” ratkaisemaan tiettyjä tehtäviä (LeCun ym. 2015). Näin ollen saattaa olla harhaanjohtavaa kuvailla kyseistä teknologiaa ”älykkyys” käsitteellä (Esposito 2022), vaikka ne selkeästi ylittävät ihmisen toimintakyvyt monissa tehtävissä, esimerkiksi suurten datamäärien analysoimisessa. Tällä hetkellä kaikki käytössä olevat tekoälyt ovat kapeita tekoälyjä. Vastakohtana kapealle tekoälylle on esitetty yleistä tekoälyä (AGI, artificial general intelligence) jolla tarkoitetaan ”aidosti älykästä” tekoälyä. Tällaisen tekoälyn on ajateltu kykenevän ratkaisemaan monimutkaisia ja laaja-alaisia ongelmia useissa eri tilanteissa ja konteksteissa. (Trajtenberg 2018) Se ei myöskään kapean tekoälyn tavoin olisi rajattu pelkästään niille alueilla, joita varten

se on kehitetty vaan se kykenisi toimimaan myös näiden alueiden tai tehtäväkuvausten ulkopuolelta. Tämänkaltainen tekoäly olisi merkittävästi lähempänä ihmisen älykkyyttä ja voi jopa ylittää sen. (Goertzel ym. 2007) Yleiseksi tekoälyksi luokiteltavia tekoälyjä ei kuitenkaan todennäköisesti tulla näkemään lähivuosina.

Käytännön kannalta kapeaa ja yleistä tekoälyä olennaisempi näkökulma on tekoälyn kehityksen laajentama käyttötapausten kirjo. Tätä kenttää ei voi tyhjentävästi etukäteen määritellä, joten olennaiseksi muodostuu kyky hahmottaa ja ohjata kokonaisuutta eri keinojen tasoilla suuntaan, joka edistää hyvän elämän ja yhteiskunnan toteutumista.

Tekoälyä ei tulisi tulkita ainoastaan irrallisena ohjelmistona tai algoritmina, vaan uusia mahdollisuuksia tarjoavana yleiskäyttöteknologiana, joka on nivoutunut osaksi laajempana sosioteknistä järjestelmää. Yhdessä erilaiset tekoälyjärjestelmät ja sovellukset muodostavat kokonaisuuden, joita hyödynnetään tukemaan päätöksentekoa ja korvaamaan ihmistyötä, lisäten eri alojen kyvykkyyttä toimia tehokkaammin tai uudella tavalla. Tämän kyvykkyyden hyödyntäminen yhteiskunnan eri osa-alueilla muuttaa yhteiskunnan toimintaa, saaden aikaan sosioteknistä muutosta. Keskittyminen tekoälyyn ainoastaan teknisenä ja laskennallisena, yhteiskunnallisesta kontekstistaan ja historiastaan erillisenä järjestelmänä saattaa näivettää keskustelua sen monista eettisistä ja poliittisista vaikutuksista sekä yhteiskunnallisista edellytyksistä (Coeckelbergh 2020; Crawford 2021). On syytä ottaa systeeminen näkökulma tekoälyn tuotantoon ja käyttötapoihin: tekoälyt nivoutuvat muiden elektronisten ja digitaalisten teknologioiden tavoin globaaleihin logistiikkaketjuihin ja infrastruktuureihin, luonnonmateriaalien kuten mineraalien louhintaan, työhön muotoihin, sekä näiden ympärille muodostuviin sosiaalisiin ja poliittisiin yhteiskunnan valtarakenteisiin (Crawford 2021). Näin ollen tekoälysääntelyn ja ohjauksen ei tulisi keskittyä vain yksittäisiin sovellutuksiin, vaan kokonaisuuteen. Esimerkiksi ihmistyön korvaaminen tekoälyratkaisulla palvelutuotannossa saattaa sisältää eettisesti ja yhteiskunnallisesti negatiivisia vaikutuksia työntekijöitä kohtaan, mutta samanaikaisesti positiivisia vaikutuksia joidenkin palvelun käyttäjäryhmien parissa ja palvelutuotannon yleisessä tehokkuudessa. Toimiva tekoälyohjaus kykenisi hahmottamaan tilannetta tavalla, jossa sekä negatiivisiin että positiivisiin vaikutuksiin kyettäisiin tarttumaan ja esimerkiksi kokonaispalvelu ja työ organisoimaan uusin tavoin.

2.4 Tekoälyn eettinen ohjaus

Yrittäessämme hahmottaa mutkikkaita ongelmia, on hyvä palata perusasioihin ja keskittyä ajattelemaan, kuinka hyvää voidaan tehdä ja edistää ja pahaa välttää ja ehkäistä silloinkin, kun ihmisten ja tekoälyjärjestelmien toiminta punoutuvat yhteen. Eettistä keskustelua tulisi käydä sekä ihmisen hyvään ja hyvinvoinnin olemukseen liittyvistä tekijöistä, että tekoälyn suhteesta yhteiskuntaan ja yhteiskunnan muutokseen. Etiikan perusperiaatteita ovat hyvän tekeminen, vahingon välttäminen, oikeudenmukaisuus ja itsemääräämisoikeus eli autonomia. Eettisten

periaatteiden nimeäminen helpottaa asioiden punnintaa, vaikka samaan periaatteeseen kuuluvat tekijät sekä eri periaatteet voivat olla vaikutuksiltaan vastakkaisia ja ristiriidassa keskenään. Vaakakupissa voivat olla ristiriidassa myös yhteiskunnan ja yksilön etu. Samoin yksilön oikeus päätöksentekoon voi sotia jonkun toisen etua vastaan.

Normit ja arvot ohjaavat tavoitteitamme ja toimintaamme, joten niillä on merkittävä rooli myös teknologian suunnittelussa ja käyttöönotossa. Arvot ovat kulttuurisesti hallitseva käsitys yksilöiden, yhteiskunnan ja ihmiskunnan hyvän elämän keskeisistä tavoitteista, ja ne ohjaavat päätöksentekoa. Yksilöllisten ja kulttuuriarvojen lisäksi on olemassa yleismaailmallisia arvoja, jotka ylittävät kulttuuriset ja kansalliset rajat, kuten YK:n ihmisoikeuksien yleismaailmallisessa julistuksessa ja Euroopan Unionin perusoikeuskirjassa vahvistetut perusarvot.

Teknologiaa on aikojen saatossa valitettavasti pidetty arvoneutraalina, kun taas etiikka on yhdistetty normatiivisiin uskomuksiin ja subjektiivisiin normeihin. Teknologioilla on kuitenkin aina omat eettiset ulottuvuutensa, jotka muokkaavat omalta osaltaan näkemyksiä ja käytäntöjä (Jasanoff 2016).

Etiikan tutkimuksessa puhutaan etiikan kolmesta osa-alueesta. Metaetiikka tutkii eettisten käsitteiden alkuperää sekä moraalisen kielen luonnetta ja pätevyyttä. Normatiivinen etiikka määrittää normeja tai standardeja. Soveltavan etiikan tavoitteena on arvojen soveltaminen eri asiayhteyksissä. Yritysmaailmassa soveltavan etiikan tutkimusalaan voi katsoa kuuluvan yritysten yhteiskuntavastuu sekä liike-elämän etiikka. Lisäksi deskriptiivisellä etiikalla on kasvava merkitys tekoälyn etiikan kannalta. Se tutkii miten moraalikäsitykset ja normit kehittyvät ja ilmenevät yksilöissä ja yhteisöissä.

Tekoälyn eettinen ohjaus on nojannut pääsääntöisesti normatiiviseen soveltavaan etiikkaan, jossa pyritään hyödyntämään eettistä harkintaa ja tietoa tunnistetuista tekoälyn eettisistä ongelmista ja mahdollisuuksista. Etiikka tarjoaa keinoja pohtia, mitä yhteinen hyvä merkitsee eri asiayhteyksissä. Eettiset periaatteet ja riskit ovat eettisistä keskusteluista syntyneitä ja niistä tiivistettyjä oppeja, jotka kehystävät pohdintaa tekoälyn ohjaus- ja koordinaatiotoimenpiteistä (esim. Jobin ym. 2019; Floridi ym. 2018). Tekoälyn vaikutuksia voidaan siis arvioida ja ohjata yleisinä eettisinä arvoina pidettyjen periaatteiden sekä tunnistettujen eettisten riskien avulla. Näin ne ohjaavat keskustelua yhteisesti jaettujen arvojen ja periaatteiden perusteella. Tekoälyn eettiset periaatteet ovat perusta tekoälyn eettiselle suunnittelulle ja eettiset haasteet antavat osviittaa siitä, minkälaisia riskejä tulisi vältellä. Nämä periaatteet auttavat luomaan kysymyksiä ja etsimään ja järjestämään vastauksia kysymyksiin. Koska tekoälyn etiikka on kuitenkin aina sidoksissa asiayhteyteen, yksikään periaatteista ei voi sellaisenaan tarjota selkeitä vastauksia kaikkiin suunnittelukysymyksiin, ja periaatteet voivat jopa olla keskenään ristiriitaisia. Käytännössä eettisiä periaatteita hyödynnetään kyseessä olevaan kontekstiin soveltuvien eettisten toimenpidesuositusten laatimiseen. Nämä kertovat ”miten käsitteellistää, analysoida ja arvioida tekoälyn suunnittelun ja soveltamisen eettisesti merkityksellisiä piirteitä sekä miten määritellä menetelmät, joilla tekoälyn suunnittelua ja käyttöä ohjataan, säännellään ja hallitaan eettisesti kestävällä tavalla.” (Hallamaa & Kalliokoski 2022)

Eettiset periaatteet auttavat myös julkisia toimijoita ja erilaisia organisaatioita hyötymään tekoälyn mahdollisuuksista ja ennakoimaan epätoivottuja sekä ymmärtämään sosiaalisesti arvostettuja toiminnan muotoja. Eettiset periaatteet luovat pohjaa julkiselle ja avoimelle keskustelulle tekoälyn tavoitteista ja hyödyntämisestä ja tämän kautta syntyvälle luottamukselle. Samalla ne toimivat indikaattorina potentiaalisista riskeistä tekoälyä hyödyntäville yrityksille (Floridi ym. 2019, 694–5).

2.5 Tekoälyn eettiset hyödyt ja haasteet

Tekoälyn eettistä keskustelua johdattavat tekoälyn potentiaalisesti merkittävät hyödyt ja haasteet (Floridi ym. 2018). Tekoälyn ennakoitu sosiotekninen kehityskaari potentiaalisten hyötyjen suhteen on laaja: eri maiden politiikkadokumenteissa tekoälystä odotetaan merkittäviä taloudellisia hyötyjä lisääntyneen tehokkuuden ja tuottavuuden vuoksi (EC 2020). Tekoälystä on povattu merkittävää tekijää hyvän tekoäly-yhteiskunnan (Wamba ym. 2021; Stahl ym. 2021; Russell ym. 2015; Makridakis 2017; Floridi & Cowls 2019; Baum 2017) ja YK:n sosiaalisten kehitystavoitteiden edistämisessä (Truby 2020; Vinuesa ym. 2020). Parhaimmillaan tekoäly tarjoaa mahdollisuuksia tukea ihmisten toimijuutta, itsensä toteuttamista, sosiaalista yhteenkuuluvuutta ja yhteiskunnan kyvykkyyttä (Floridi ym. 2018). Tekoälyn arvellaan voivan tukea ihmisten ja yhteiskunnan hyvinvointia ja yhteistä hyvää, eli se voi parhaimmillaan olla ihmiskunnan kukoistusta tukeva teknologia (Stahl 2021; AI HLEG 2019; 4). Tällaisten odotusten ja visioitujen mahdollisuuksien vuoksi tekoälyn alikäyttöä pidetään eräänlaisena riskinä: tekoälyn oikeanlaisen potentiaalin hyödyntämättä jättäminen on eettisesti arveluttavaa (Floridi 2018). Toisaalta riskinä on myös tekoälyn väärinkäyttö ja sen hyödyntäminen tarkoituksiin, jotka eivät tue ihmiskunnan kukoistusta pitkällä aikavälillä (Bostrom 2014).

Tekoälyn eettiset haasteet voidaan karkeasti ottaen jakaa kolmeen luokkaan (Stahl 2021, katso taulukko 1): Ensimmäinen liittyy teknisellä tasolla koneoppimisesta ja algoritmeista juontuviin haasteisiin. Tällaisia ovat esimerkiksi datan vääristyneisyys ja diskriminaatio, datan transparenssi ja validointi, datan turvaaminen ja turvallisuus. Toinen ovat emergentit yhteiskunnalliset haasteet, kuten muutokset ihmisen autonomiassa, ihmisen ja koneen välisen vuorovaikutuksen muutos, luottamus tekoälyjärjestelmiin ja niiden sosiaalinen hyväksyntä, tasa-arvokysymykset ja yhteiskunnalliset valtarakenteet, epäoikeudenmukaisuus ja sorto, sekä sodankäyntiin, työn muutokseen ja ekologiseen kestävyyteen liittyvät kysymykset. Kolmas luokka viittaa tulevaisuuden metafyysisiin kysymyksiin potentiaalisen superälykkään tekoälyn kehittymisestä. Näissä keskusteluissa tekoälyn muodostamaa riskiä on verrattu ihmiskuntaa uhkaaviin tekijöihin, kuten ilmastonmuutokseen, ennakoidun superälykkään toimijan ylivaltaan ja ihmiskunnan tuhoon (Bostrom 2014).

Taulukko 1. Näkökulmia tekoälyn eettisiin haasteisiin (muokattu Stahl 2021)

Tekninen näkökulma	Episteemiset kysymykset: Luotettavuus, vääristymät, väärät johtopäätökset, läpinäkyvyys, mustat laatikot. Vastuuvelvollisuus, yksityisyyden suoja, tietosuoja, turvallisuus.
Sosiotekninen näkökulma	Taloudelliset vaikutukset, oikeudenmukaisuus, autonomia ja vapaus, epätasa-arvo, luottamus, ympäristövaikutukset, sosiaalinen kestävyys, autonominen sodankäynti, rikollinen käyttö
Pitkän aikavälin vaikutukset	Ekologinen, sosiaalinen ja taloudellinen kestävyys, superälykäs tekoäly, ihmisautonomian kaventuminen, julkisen ohjauksen ja hallinnan edellytysten muuttuminen

Mittlestadt (2016) ja Tsamados (2021) kumppaneineen ovat tutkimuksissaan ryhmitelleet tekoälyalgoritmien keskeiset eettiset haasteet kolmeen luokkaan:

1. episteemiset kysymykset siitä, miten algoritmi hyödyntää dataa tuottaakseen tietyn lopputuloksen tai johtopäätöksen,
2. normatiiviset kysymykset siitä, miten näitä johtopäätöksiä hyödynnetään ohjaamaan päätöksentekoa ja toimintaa,
3. algoritmien hyödyntämisen vastuullisuuteen ja vastuuvelvollisuuteen liittyvät kysymykset.

Algoritmien **episteemiset kysymykset**, eli tietoa ja tiedontuotantoa koskevat kysymykset voidaan jakaa kolmeen ryhmään. Ensinnäkin algoritmin tekemät johtopäätökset perustuvat todennäköisyyslaskelmiin, eikä algoritmin hyödyntämä tietopohja ole täydellinen, mikä voi johtaa vääriin johtopäätöksiin tai tuloksiin. Toiseksi algoritmit ovat usein niin sanottuja ’mustia laatikoita’; niiden toimintaa on niiden vaikeaselkoisuuden läpinäkymättömyyden vuoksi hankala tai miltei mahdotonta tutkia, ohjata ja parannella. Kolmanneksi algoritmit tekevät johtopäätöksensä niihin syötetyn datan perusteella, mikä itsessään saattaa olla vääristynyttä, jolloin myös algoritmin tulokset vääristyvät. Algoritmien tuottaman tiedon **normatiivisia kysymyksiä** on kahtalaisia. Yhtäältä ne saattavat olla epäoikeudenmukaisia hyödyttämällä yhtä ihmisryhmää enemmän kuin muita. Näin on käynyt esimerkiksi Yhdysvaltojen poliisien käyttämän ennakoivan toiminnanohjausohjelman kanssa, jonka epäillään johtaneen tummaihoisten ihmisten suhteettoman suureen pidätysmäärään (Selbst 2017). Toiseksi tällaisten algoritmien hyödyntäminen johtaa kysymyksiin yksityisyyden suojasta ja ihmisautonomiasta. Esimerkiksi digijättien keräämä käyttäytymisdata ja näihin liitetyt mainonnan kohdentamisen algoritmit herättävät kysymyksiä siitä, missä määrin algoritmien on sopivaa vaikuttaa ihmismielen tiedostomattomaan toimintaan, ja minkälaisia negatiivisia vaikutuksia tästä syntyy (Botes 2022). Kolmantena algoritmien eettisiä haasteita kuvaavana luokkana on **vastuuvelvollisuus** (accountability). Kun algoritmi aiheuttaa ongelmia, on tästä moraalisesti vastuullista tahoa vaikea määritellä. Näin on käynyt esimerkiksi autonomisten autojen aiheuttamien kuolemaan johtaneiden onnettomuuksien jälkipuinnissa.

2.6 Eettisen ja vastuullisen tekoälyn periaatteet

Eri valtioiden, suuryritysten ja kansainvälisten järjestöjen tekoälyn etiikkaa linjaavissa dokumenteissa on yhtäläisyyksiä. Laajan, Jobin ja kumppaneiden (2019) toteuttaman 84 eettisen asiakirjan sisällönanalyysin perusteella dokumentit näyttäisivät viittaavaan viiteen periaatteeseen. Näitä ovat läpinäkyvyys, oikeudenmukaisuus, vahingonteon välttäminen, vastuullisuus ja yksityisyys (Jobin ym. 2019), joista läpinäkyvyysperiaate (eng. transparency) oli yleisin. Euroopan unionissa vaatimuksina on näiden lisäksi tekoälytoiminnan selitettävyys ja ihmislähtöisyys (esim. AI HLEG 2019). EU maiden poliittisissa dokumenteissa

vaaditaan lisää luottamusta, vastuullisuutta, vastuuvelvollisuutta, läpinäkyvyyttä ja turvallisuutta tekoälyn kehitykseen ja hyödyntämiseen (Van Roy ym. 2021) AI4People-hanke on kartoittanut yleismaailmallisia tekoälyn eettisiä periaatteita tutkimalla useiden kansainvälisesti merkittävien organisaatioiden laatimia eettisiä periaatteita. Hankkeen (Floridi ym. 2018) mukaan tekoälyn suunnittelua ja käyttöä ohjaavia periaatteita on viisi: hyvän tekeminen (beneficience), vahingon välttäminen (non-maleficience), itsemääräämisoikeus (autonomy), oikeudenmukaisuus (justice) ja selitettävyys (explicability).

Taulukko 2. Eettisen ja vastuullisen tekoälyn käytön periaatteita (muokattu Mikaleff ym. 2022, 259).

Oikeudenmukaisuus	Tekoälyjärjestelmien tulisi olla osallistavia ja monimuotoisia, eikä niiden tule olla syrjiviä.
Vastuullisuus	Tekoälyjärjestelmien tulisi olla avoimia ja läpinäkyviä prosessien ja tulosten osalta. Niiden tulisi tukea jäljitettävyyttä, selitettävyyttä ja käyttäjäviestintää.
Robustisuus ja turvallisuus	Tekoälyjärjestelmiä olisi kehitettävä huomioiden vastuullisuus, vastuuvelvollisuus sekä eettiset periaatteet.
Tiedonhallinta	Tiedonhallinnassa tulisi huomioida datan laatu ja eheys koko tekoälyjärjestelmän elinkaaren ajan.
Lait ja asetukset	Tekoälyjärjestelmien tulisi noudattaa niiden toimintaa sääteleviä lakeja ja asetuksia.
Valvonta	Tekoälyjärjestelmien olisi tuotettava konkreettista hyötyä ihmisille ja niiden tulisi olla ihmisen valvonnassa.
Yhteiskunnan ja ympäristön hyvinvointi	Tekoälyjärjestelmien tulisi edistää kestävyyttä sekä ekologista ja sosiaalista vastuullisuutta eikä niistä saisi aiheutua haittaa.

Näiden periaatteiden huomioiminen tekoälyjärjestelmien kehitystyössä merkitsee, että yleisesti ottaen tekoälyn on parannettava yksilöllistä ja kollektiivista hyvinvointia. Kehitystyössä tulee kunnioittaa yhtäläisesti kaikkien ihmisten ihmisarvoa. Tekoälyjärjestelmät eivät voi tuottaa epäoikeudenmukaisesti puolueellisia tai syrjiviä tuotoksia eivätkä heikentää demokraattisia prosesseja tai ihmisten omaa harkintaa ja dialogia. Ihmiskeskeisen suunnittelun periaatteita noudattaen järjestelmien on tarjottava mahdollisuus ihmisen valinnoille sekä ihmisen valvonnan ja määräysvallan varmistamiselle tekoälyjärjestelmien työprosesseissa. Tekoälyjärjestelmät eivät saa perusteettomasti pakottaa, alistaa, johtaa harhaan, manipuloida, ehdollistaa tai holhota ihmisiä, vaan järjestelmien kanssa toimivien ihmisten on voitava säilyttää täysimääräinen ja tehokas itsemääräämisoikeus. Myös haavoittuvassa asemassa olevat tulee huomioida järjestelmien kehittämisessä ja käyttöönotossa. Erityisesti tulee kiinnittää huomiota tilanteisiin, joissa vallitsee vallan tai tiedon epätasapaino (esimerkiksi (työnantaja/työntekijä; yritykset/kuluttajat; päättäjät/kansalaiset). Tekoälyn

toimintaperiaatteiden (algoritmien) tulee olla läpinäkyviä, selitettävissä ja jäljitettävissä.

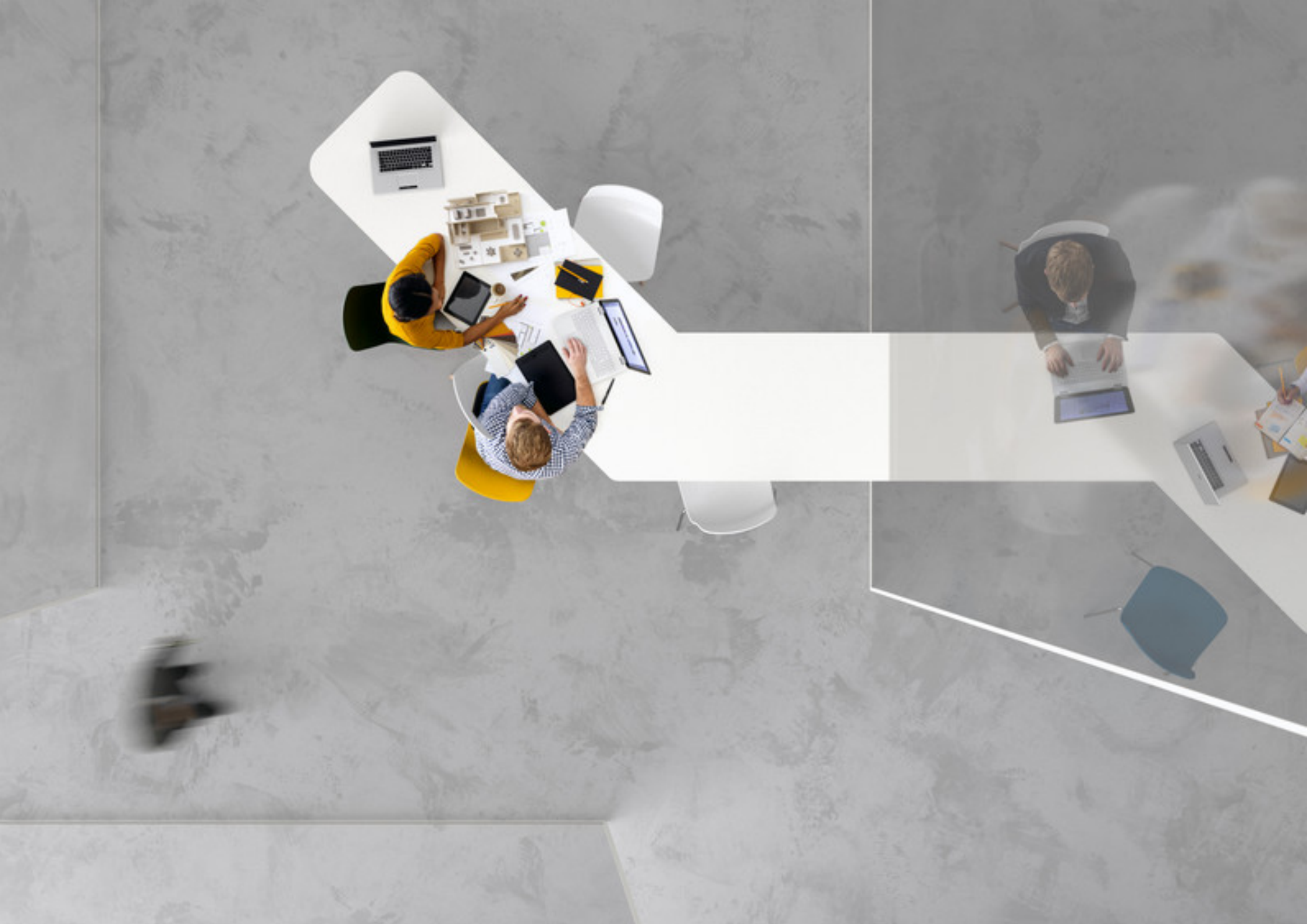
Eettisten periaatteiden toteutuminen käytännössä riippuu pitkälti niiden omaksumisesta valtioiden, kansainvälisten yhteisöjen, sekä erilaisten organisaatioiden ja yritysten ohjausmekanismeissa. Nykyisellään tekoälyä ohjaavia eettisiä suosituksia, ohjeistuksia ja periaatteita on kritisoitu siitä, ettei niillä ei näytä olevan käytännön vaikuttavuutta (esim. Larsson 2020). Eettisiä periaatteita on kuvailtu ”mielipiteiden ja kontribuutioiden kakofoniaksi” (Stahl ym. 2021), jotka eivät näytä juurtuvan politiikkatoimenpiteisiin eikä yritysten käytäntöihin, mahdollistaen niin sanottua eettistä viherpesua yrityksissä (Fjeld ym. 2020; Floridi 2019; Mittelstadt ym. 2019; Morley ym. 2019; Bietti 2021). Tutkimuskenttä on myös kehittynyt nopeasti. Nykyisin on olemassa lukuisia erilaisia ohjeistuksia ja käytäntöjä yhteistä hyvää tavoittelevan tekoälyn suunnitteluun (Floridi ym. 2020) sekä konkreettisia työkaluja eettisten periaatteiden toteuttamiseksi koneoppimista kehittäville tutkijoille ja koodareille (Morley ym. 2020). Tutkijat ovat niin ikään laatineet monia erilaisia enemmän tai vähemmän suoraan toteutettavissa olevia ehdotuksia eettisten periaatteiden juurruttamiseksi tekoälyn ohjaus- ja koordinaatiomekanismeihin (esim. Stix ym. 2021; Floridi ym. 2018; Tsamados ym. 2021; Donahoe & Metzger 2019; Yeung ym. 2019). Niistä osa on soveltaen kirjattu osaksi EU:n tekoälyn ohjauksen toimenpidesuosituksia (AI HLEG 2019) sekä OECD:n periaatteisiin (OECD 2019).

2.7 Ohjauksen kehystäminen: tavoitteet, hyödyt ja haasteet eettisten ohjauskeinojen muodostamisen perustana

Tekoälyyn liittyvät mahdolliset ongelmat ja eettiset kysymykset ovat monimutkaisia ja vaikuttavat koko yhteiskuntaan (esim. Floridi ym. 2018; Tsamados ym. 2021; Coeckelbergh 2020; Crawford 2021; Zuboff 2019). Tekoälyyn liittyvien riskien, mahdollisuuksien ja tavoitteiden määritelmät ohjaavat keskustelua siitä, minkälaisia toimenpiteitä tulisi soveltaa tekoälyn ohjaamiseksi erilaisissa asiayhteyksissä. Kuten kappaleessa 3.2 mainitsimme, tämän raportin näkökulma on, että tekoälyä ei tulisi tulkita vain erillisenä ohjelmistona (tai algoritmeina), vaan yleiskäyttöisenä teknologiana, joka on osa laajempaa sosioteknistä järjestelmää. Keskittyminen tekoälyyn vain teknisenä ja laskennallisena järjestelmänä, erillään sen yhteiskunnallisesta kontekstista ja historiasta, kaventaisi suhteettomasti keskustelua sen eettisistä vaikutuksista, yhteiskunnallisista edellytyksistä ja mahdollisesta yhteiskunnallisesta muutoksesta (Coeckelbergh 2020; Crawford 2021).

Tekoälyn ohjauskeinojen kehittäminen nojaa odotuksiin ja tulevaisuudenkuviin tekoälyn riskeistä, potentiaalista ja roolista yhteiskunnassa. Puhutaan tekoälyn kehystämisestä: tulevaisuusvisiot ja odotukset ovat keskeisiä sille, miten tekoälyyn liittyvät riskit ja mahdollisuudet ymmärretään ja huomioidaan yritysten keskuudessa sekä poliittisissa asiakirjoissa ja vastaavasti millaisia ohjauskeinoja näiden perusteella on syytä kehittää. Visiot ja odotukset riskeistä ja mahdollisuuksista ovat myös performatiivisia sillä ne kehystävät ajattelua siitä, millaiset ohjauskeinot ovat tarpeellisia, merkityksellisiä tai mahdollisia (Konrad & Böhle 2019; Mager & Katzenbach 2021). Samalla ne toimivat perustana tekoälyn rahoitukselle, tekoälyn kehittämisen tavoitteille ja varsinaiselle ohjaukselle (Ulnicane ym. 2020; Bareis & Katzenbach 2021).

Pohjimmiltaan kehystäminen riippuu siitä, missä määrin eri sidosryhmien näkemykset otetaan huomioon kollektiivisia sosioteknisiä tulevaisuuskuvauksia ja odotuksia muodostettaessa. Radu (2021) on tutkimuksissaan havainnut, että kansallisten tekoälystrategioiden sisältöjä ovat muodostaneet pääosin epämääräisin perustein valitut tutkijoiden, hallinnon alojen ja elinkeinoelämän edustajat. Nämä toimijat muovaavat tekoälyn tavoitteita ja vaikuttavat yhdessä politiikkatoimenpiteisiin ja tekoälyn ohjauksen saamiin sisältöihin, jotka lopulta saattavat hyödyttää tiettyjä toimijoita muita enemmän tai näyttäytyä kyseenalaisilta yhteisen hyvän, vastuullisuuden ja eettisten periaatteiden valossa. Poliittisten toimenpiteiden ja kansallisten strategioiden kehystämisen näkökulmasta olisikin tärkeää välttää yksittäisten eturyhmien vaikutusvaltaa tekoälyä koskevassa politiikanteossa ja käydä mahdollisimman moniäänistä ja empiiriseen näyttöön perustuvaa sidosryhmäkeskustelua tekoälyn käytöstä ja haasteista erilaisissa asiayhteyksissä (Sun 2019; Delacroix 2021).



3 Ehdotuksia tekoälyn ohjauksen kehittämiseen: periaatteet ja keinovalikoima

Esittelemme seuraavaksi tutkimuskirjallisuudessa esiintyviä tekoälyn julkisen ohjauksen kehitysehdotuksia. Ehdotukset perustuvat Sigfrids ja kumppaneiden (2022) toteuttamaan systemaattiseen kirjallisuuskatsaukseen, jonka menetelmät on kuvattu liitteessä 1. Katsauksessa hyödynsimme temaattista analyysiä selvittämään tekoälyn julkisen ohjauksen kehittämistä käsittelevässä 21 tutkimusartikkeleissa olevia ongelmanmääritelmiä, periaatteita ja tavoitteita, sekä keinoja ja menettelyitä, joita tarvitaan tekoälyn ohjauksen kehittämiseksi.

Luokittelimme kehitysehdotukset neljään teemaan: (1) Hallintamallit ohjauskeinojen kehittäjien tukena, (2) Ohjauksen- ja koordinaation menettelyt, (3) Eettisten periaatteiden juurruttamisen keinot, (4) Tarvittavat institutionaaliset mekanismit. Koonnin tavoitteena on herättää keskustelua tekoälyn julkisen ohjauksen kehittämisestä sekä laajentaa keinovalikoimaa, jolla julkisen hallinnon on mahdollista avittaa ihmisoikeuksien, etiikan ja vastuullisuuden toteuttamista tekoälyä koskevassa poliittisessa päätöksenteossa sekä organisaatioiden käytännöissä.

3.1 Ohjauksen ja koordinaation tulee perustua kokonaisvaltaiseen näkemykseen tekoälyilmiöstä

Ensimmäinen teema tarkastelee niin sanottuja hallintamalleja (eng. governance model tai framework) keinona muodostaa kokonaisvaltainen lähestymistapa tekoälyn ohjauksen kehittämiseen julkishallinnossa. Koska tekoälyilmiö on laaja-alainen, nopeasti kehittyvä ja sen eettiset ongelmat kontekstisidonnaisia, täytyy tekoälyilmiö ohjauskeinoja kehittäessä huomioida mahdollisimman kokonaisvaltaisesti, jotta vältetään ongelman määrittelyiden ja niiden perusteella tehtävän päätöksenteon yksipuolisuus. Tekoälyn ohjauksessa on tärkeää sovittaa yhdessä sovittuja sääntöjä ja eettisiä periaatteita paikallisiin asiayhteyksiin ja niistä kumpuaviin erilliskysymyksiin, toimijoiden intresseihin ja tarpeisiin yhdessä päätöksentekijöiden ja käyttäjien kanssa. Hallintamallit ovat heuristisia kuvauksia siitä, miten tekoälyn haasteisiin ja mahdollisuuksiin voidaan vastata yhteiskunnan eri tasoilla ja minkälaisia näkökulmia, prosesseja ja toimenpiteitä tehtävän

toteuttamiseksi tulisi huomioida. Hallintamallien tavoitteena on auttaa tutkijoita ja julkishallinnon edustajia pohtimaan tarvittavia institutionaalisia mekanismeja ja keinoja, joiden kautta tekoälyilmiö voidaan käsitteellistää hallittavana ilmiönä mahdollisimman kokonaisvaltaisesti, mutta paikalliseen asiayhteyteen soveltuen. Sen sijaan, että keskityttäisiin yksittäisiin tekoälyn eettisiin haasteisiin (kuten yksityisyyden suojaan, turvallisuuteen tai oikeudenmukaisuuteen), mallit pyrkivät mahdollistamaan sellaisen sääntelyn ja kollektiivisen päätöksenteon muodon, joka perustuu ymmärrykseen niin tekoälyilmiön kontekstisidonnaisista ilmentymistä, kuin sen sosioteknisistä, eettisistä ja lainsäädännöllisistä vaikutuksista. Toisin sanoen, mallit tarjoavat "käsitteellisen linssin, jonka avulla yhteiskunnat voivat ajatella kollektiivisesti ja tehdä tietoon perustuvia poliittisia päätöksiä siitä, mitä, milloin ja miten tekoälyn käyttöä ja sovelluksia olisi säänneltävä" (de Almeida ym. 2021, 505).

Yleiskuva analysoiduissa teksteissä on, että tekoälyn ohjaus- ja koordinaatio ovat mekanismeja, joilla suojellaan yleistä etua, minimoidaan riskejä ja tasapainotetaan yhteiskunnan eri sidosryhmien etuja (Baldwin, 2012), samalla varmistaen ihmisarvojen toteutuminen ohjauksen perustana (Cath ym. 2018; Yeung ym. 2019; Gasser & Almeida 2017). Tekoälyn hallintamallin kannalta tämä tarkoittaa, että mallin tulisi tukea tekoälyn kollektiivisen ymmärryksen lisäämistä eri yhteyksissä, helpottaa konsensuksen rakentamista eri sidosryhmien välillä sekä tukea arvojen ja etujen välisiä kustannus-hyötyanalyysejä (Gasser & Almeida 2017; Wirtz ym. 2018; Rahwan 2018; Wallach & Marchant 2018; Yeung ym. 2019). Ohjauskeinojen kehittämisen olisi perustuttava asiaankuuluvien sidosryhmien ja kansalaisten näkemyksiin sekä kollektiiviseen ymmärrykseen tekoälyn yhteiskunnallisista hyödyistä (Cath ym. 2018; Wirtz ym. 2020). Lisäksi tekoälyn kehittäjien tulisi olla riippumattoman sääntelyviranomaisen lakisääteisen valvonnan alaisia (de Almeida ym. 2021; Wallach & Marchant 2018; Yeung ym. 2019; Rahwan 2018).

Hallintamallien olisi niin ikään tuettava sellaisen kokonaisvaltaisen ja integroidun ohjauskeinovalikoiman kehittämistä, jossa tekoälyn ohjauksen eri ulottuvuudet lainsäädännöstä eettisiin periaatteisiin ja teknisiin ratkaisuihin nähdään systeemisenä kokonaisuutena (Gasser & Almeida 2017). Hallintamalleissa yhdistyvät tiedon tuottaminen tekoälyn kontekstisidonnaisista ilmiöistä, sosioteknisistä, eettisistä ja lainsäädännöllisistä vaikutuksista sekä erilaisten sidosryhmien ja kansalaisten näkemykset, kollektiivisiin päätöksenteon ja lainsäädännön kehittämisen muotoihin (Gasser & Almeida 2017; Wirtz ym. 2018; de Almeida ym. 2021; Yeung ym. 2019). Wirtzin ym. (2020) sekä Gasser ja Almeidan (2017) ehdotuksien mukaan hallintamallin tulisi huomioida:

- ***Tekoälyteknologioiden ja -palveluiden monimuotoisuus***. Tekoälyteknologioiden erilaiset käyttökontekstit edellyttävät toisistaan hyvin erilaisia ohjaus- ja sääntelykeinoja, joita tulisi huomioida niin teknisten, organisaatiokohtaisten kuin poliittisten ohjaustoimenpiteiden tasolla.
- ***Tekoälyn vaikutukset***. Ohjaustoimenpiteiden tulisi perustua laaja-alaiseen ymmärrykseen tekoälyn aiheuttamista yhteiskunnallisista, eettisistä ja lainsäädännöllisistä suorista ja välillisistä vaikutuksista.

- ***Tekoälyn sääntely.*** Tehtävänä on määritellä, minkälaisiin tekoälyn haasteisiin sääntelyllä tulisi vastata ja minkälaisia vastuualueita ja instituutioita tekoälyn sääntelyn liittyy.
- ***Julkisten ohjauskeinojen toimeenpano***. Ohjausmekanismit voivat sisältää kovaa ja pehmeää sääntelyä, kuten yhteiskunnallisia ja oikeudellisia normeja, eettisiä periaatteita ja käytännesääntöjä, sääntelyä ja lainsäädäntöä, sekä teknisiä ja organisaatioiden käytäntöjä, kuten tiedonhallintatyökaluja, standardeja ja sertifiointeja. Politiikkatoimenpiteiden toimeenpanossa olisi huomioitava tekoälyteknologioiden ja palveluiden erilaiset kontekstit ja niihin liittyvät ohjaustoimenpiteet niin teknisellä, organisaatiokohtaisella kuin politiikkatoimenpiteiden tasolla.
- ***Yhteistoiminnallisuuden mahdollistaminen***. Sidosryhmien erilaisten tavoitteiden ja ristiriitaisten etujen tasapainottamiseksi on tärkeää rakentaa luottamusta, yhteisiä arvoja ja osallistumisen motivaatiota sidosryhmien välille (Wirtz ym. 2020, 825). Toiminnan tulisi helpottaa konsensuksen rakentamista eri sidosryhmien välillä ja tukea tekoälyn hyödyntämisen kustannus-hyötyanalyysejä arvojen ja eri sidosryhmien etujen välillä.

On huomioitava, että ehdotetut hallintamallit ovat teoreettisia malleja, joita ei ole kokeiltu empiirisesti eikä niitä ole kehitetty yhdessä sidosryhmien kanssa. Ne ovat alustavia ehdotuksia keskeisistä asioista, jotka tulisi huomioida tekoälyn hyvän julkisen ohjauksen ja koordinoinnin kehittämisen perustana.

3.2 Ohjauksen ja koordinaation menettelyiden tulee olla osallistavia, ketteriä ja mukautuvia

Toinen teema käsittelee ohjauksen ja koordinaation menettelyiden kehittämistä hallinnon joustavuuden eli ketteryyden ja mukautuvuuden lisäämiseksi. Vallitseva näkemys on, että tekoäly yhteiskunnallisena ilmiönä on monimutkainen, laaja-alainen ja nopeasti kehittyvä, minkä vuoksi hierarkkisia ohjausmuotoja olisi täydennettävä ketterillä ja mukautuvilla ohjauskeinoilla. Teemassa käsitellään kysymystä siitä, miten julkista ohjausta ja päätöksentekoa voidaan parantaa huomioimalla tekoälyn käytön ja vaikutusten kontekstisidonnaisuus ja muuttuva luonne, samalla välttäen uuden ja nopeasti kehittyvän teknologian ohjaukseen liittyvät haasteet. Ehdotetuissa ratkaisuissa korostetaan kolmea lähestymistapaa:

(1) ketterien ja mukautuvien ohjauskeinojen hyödyntäminen, (2) yhteissääntelyn hyödyntäminen tekoälyn sääntelyn kehittämisessä, ja (3) jatkuvan teknologisen muutoksen huomioiminen ohjauksen kehittämisessä (ks. Taulukko 3).

Ketterä ja mukautuva ohjaus ja koordinaatio (eng. adaptive governance, agile governance) ovat lähekkäisiä käsitteitä, jotka perustuvat ajatukseen ohjauskeinojen jatkuvasta mukautumisesta yhteiskunnalliseen ja teknologiseen muutokseen: tarvitaan yleisiä periaatteita ohjaamaan toimintaa, mutta tarpeeksi joustoa soveltaa ohjauskeinoja tilanteen muuttuessa. Näkemyksen mukaan julkishallinnon tulisi tukea sidosryhmien vaikutusvallan ja osallisuuden lisäämistä päätöksenteossa. Julkishallinnon tulisi niin ikään tukea erilaisten sidosryhmien näkemysten,

intressien, tarpeiden ja etujen joustavaa yhteensovittamista (Winfield & Jirotka 2018, Wallach & Marchant 2018; Gasser & Almeida 2018; Sun & Medaglia 2019; Ulnicane ym. 2021).

Mukautuvan ohjauksen tavoitteena on varmistaa julkishallinnon ajantasainen ja laaja-alainen ymmärrys tekoälyn vaikutuksista ja sen kehityksestä sekä sopeutuminen muuttuviin olosuhteisiin. Tämä merkitsee hajautettua, alhaalta ylöspäin suuntautuvaa ja osallistavaa päätöksentekoa sekä hallinnolle sisäisen ja ulkoisen asiantuntemuksen hyödyntämistä päätöksenteon tukena (Sun & Medaglia 2019, Janssen & van der Voort 2016). **Ketterillä ohjauskeinoilla puolestaan pyritään varmistamaan, että innovaatiotoiminta on yleisen edun mukaista tilanteessa, jossa teknologinen muutos on nopeaa.** Muutokseen mukautuminen edellyttää laaja-alaisesti eri sidosryhmien näkökulmien huomioimista eettisten ja vastuullisuusperiaatteiden ohella. Keskeistä ketterässä ohjauksessa ovat niin sanotut pehmeät ohjauskeinot, kuten informaatio-ohjaus, eettiset periaatteet ja standardit, joiden toimeenpanoa julkishallinnon olisi tuettava (Winfield & Jirotka 2018, Wallach & Marchant 2018).

Taulukko 3. Menettelyt ohjauksen ja koordinaation kehittämiseksi

Näkökulma	Kuvaus	Viitteet
Ketterä ja mukautuva ohjaus- ja koordinaatio	• Ohjauskeinojen tulisi perustua ajantasaiseen ja laaja-alaiseen ymmärrykseen tekoälyn vaikutuksista ja sen kehityksestä • Tekoälyn julkisen ohjauksen ja koordinaation tulisi tukea sidosryhmien osallisuuden lisäämistä päätöksenteossa. • Tekoälyn julkisen ohjauksen ja koordinaation tulisi tukea erilaisten sidosryhmien näkemysten, intressien, tarpeiden ja etujen joustavaa yhteensovittamista tekoälykysymyksissä • Ohjauskeinojen kehittämisessä tulisi hyödyntää hajautettua päätöksentekoa sekä hallinnolle sisäisen että ulkoisen asiantuntemuksen hyödyntämistä • Tekoälyn tavoitteista ja niitä ohjaavista periaatteista ja säännöistä olisi käytävä kriittistä julkista keskustelua • Mainittuja tehtäviä varten tulisi perustaa tekoälykysymyksiä koordinoiva komitea • Ohjauskeinoja ohjaavina normatiivisina periaatteina tulisi eettisten periaatteiden lisäksi olla vastuullisen tutkimuksen ja innovoinnin (RRI) periaatteet	Ulnicane ym. 2021; Winfield & Jirotka 2018; Wallach & Marchant 2018; Sun & Medaglia 2019; Clarke 2019; Buhmann & Fieseler 2021; Almeida ym. 2021; Floridi ym. 2018, Cath ym. 2018
Tekoälyn yhteissääntely	• Tekoälyn sääntelyn tulisi perustua niin sanottuun yhteissääntelyyn, jossa teollisuuden ja muiden sidosryhmien edustajat neuvottelevat yhdessä julkishallinnon kanssa lakisääteisistä velvoitteista.	Clarke 2019

Pitkän aikavälin tekoälyä ohjaavat strategiat	• Ohjauskeinojen olisi varmistettava innovaatiotoiminnan yleisen edun mukaisuus myös tilanteessa, jossa teknologinen muutos on nopeaa • Tulisi laatia keskipitkän ja pitkän aikavälin tekoälyä koskeva strategia, jonka avulla ohjauskeinoja voidaan ketterästi sovittaa yhteiskunnalliseen muutokseen. • Nykyisin teknologian ohjaus keskittyy nykyisten ohjauskeinojen soveltamiseen ja kehittämiseen. Tämän lisäksi tarvitaan tutkimusta siitä, miten hyvän hallinnon ja ohjauksen operationalisoinnin edellytykset saattavat muuttua tekoälykehityksen johdosta.	Liu & Maas 2021

Ohjauskeinoja tulisi kehittää siten, että ne ovat yhteensovittavissa teknologiseen muutokseen myös pitkällä aikavälillä. Ohjauskeinoja kehittäessä tulee ennakoiden huomioida tekoälyn pitkän aikavälin haasteet ja ongelmat, etenkin sellaiset, jotka muuttavat toivotunlaisen ohjauksen ja koordinaation edellytyksiä (Liu & Maas 2021). Tällainen uhkakuva saattaa olla esimerkiksi sosiaalisen median algoritmien tuottaman poliittisen polarisaation ja yhteisymmärryksen puute, joka saattaa heikentää demokraattista päätöksenteon edellytyksiä (Nemitz 2018; Silverman ym 2020). Liu ja Maas (2021) korostavat, että nykyisillä ohjauskeinoilla on heikot mahdollisuudet sopeutua nopeatempoisten teknologiainnovaatioiden aiheuttamiin muutoksiin ja siten turvata tekoälyn ohjausta koskevia pitkän aikavälin strategioita. Vaikka julkishallinnossa keskitytään politiikkatoimenpiteiden ja ohjauskeinojen soveltamiseen ja kehittämiseen, konkreettisten toimenpiteiden yhdistäminen pitkän aikavälin ohjausstrategiaan edellyttää ennakoivaa ja tulevaisuuteen suuntautunutta tutkimusta. Liu ja Maas (2021) ehdottavat osaratkaisuksi ”ongelmanetsintämenetelmää”, eli tieteellistä tutkimusta siitä, miten ohjauksen keskeisten periaatteiden toteuttamisen edellytykset ja mahdollisuudet saattavat ajan myötä muuttua ja miten muutokseen tulisi varautua.

Mukautuvaa ja ketterää ohjausta ja koordinaatiota ehdottavat kirjoittajat puoltavat vastuullisen tutkimuksen ja innovoinnin (RRI) periaatteiden soveltamista tekoälyn ohjauskeinojen kehittämisen perustana (Winfield & Jirotka 2018; Ulnicane ym. 2021). RRI:n tavoitteena on sovittaa tutkimus- ja kehitystoiminnan prosessit ja tulokset yhteen yhteiskunnallisten arvojen, tarpeiden ja odotusten kanssa. RRI-periaatteisiin kuuluvat **ennakointi** (innovaatiotoiminnan sosiaalisten, taloudellisten ja ympäristövaikutusten analysointi), **refleksiivisyys** (innovaation taustalla olevien motiivien ja tarkoitusten pohtiminen), **osallisuus** (erilaisten sidosryhmien ja kansalaisten etujen, arvojen ja näkökulmien nostaminen yhteiseen keskusteluun) ja **responsiivisuus** (oppiminen sekä tavoitteiden ja operatiivisten käytäntöjen muuttaminen tarvittaessa) (Owen ym. 2013). RRI voi tukea eettistä tekoälyn ohjausta erityisesti korostamalla, että tekoälyn eettisiä kysymyksiä tulee käsitellä ennakoivasti ja reflektiivisesti.

Buhmann ja Fieseler (2021) katsovat, että RRI:n toteutumiseksi tekoälyn ohjauksessa, tekoälyn tavoitteista ja niitä ohjaavista periaatteista ja säännöistä olisi käytävä kriittistä julkista keskustelua, joka heijastaisi "tavallisen kansalaisen"

voimaantunutta ääntä ja näkökulmaa. Tekoälyn etiikkaa ja vastuullisuutta koskevaa vuoropuhelua haastavat kuitenkin tiedon epäsymmetriat. Hyvän kansalaiskeskustelun ja dialogin mahdollistamiseksi olisi tuettava kansalaisten osaamista ja ymmärrystä tekoälyilmiöstä, sen eettisistä haasteista ja tavoitteista. Tämä edellyttää tiettyjen ”kommunikatiivisten periaatteiden” toteutumista (Buhmann & Fieseler 2021). Näitä ovat a) avoimet foorumit, joissa jokainen toimija voi osallistua keskusteluun, b) keskustelijoiden ymmärrys käsiteltävästä aiheesta tulisi olla mahdollisimman hyvä, joten sitä pitää tukea, c) kaikki argumentit tulisi ottaa huomioon, jotta asiaa voidaan arvioida mahdollisimman monesta näkökulmasta, ja d) erilaisten ehdotusten ja huolenaiheiden pitäisi voida vaikuttaa lopullisiin suosituksiin ja päätöksentekoon.

Clarke (2019) kyseenalaistaa pehmeiden ohjauskeinojen merkityksen tekoälyn ohjauksessa (vrt. Wallach & Marchant 2018). Hänen mukaansa ne ovat itsesääntelyssä tärkeitä, mutta tehottomia, jos toimintaa on tarkoitus ohjata yhteiskunnallisten tavoitteiden mukaisesti. Tutkimuksessaan tekoälyn erilaisista sääntelyä koskevista vaihtoehdoista, Clarke päätyy ehdottamaan **yhteissääntelyä** parhaana tapana edistää ketterää, eri osapuolten näkökulmasta mielekästä ja tehokasta sääntelyä. Yhteissääntelyllä tarkoitetaan mallia, jossa teollisuus, sidosryhmät ja viranomaiset neuvottelevat yhdessä oikeudellisista velvoitteista. Tuloksena ovat täytäntöönpanokelpoiset säännöt. Prosessissa on otettava huomioon kaikkien osapuolten tarpeet, ja on varmistettava, että institutionaaliset tai markkinavoimat eivät vääristä keskustelua. Clarke (2019) tarjoaa konkreettiset puitteet tällaisen sääntelyjärjestelmän suunnittelulle ja arvioinnille. Hän esittää tarkoitusta varten käsitteellisen kehyksen, joka arvioi sääntelyprosessin avoimuutta, sidosryhmien etujen huomioimista, sääntelymekanismien artikulointia ja täytäntöönpanoa sekä vastuuvelvollisuutta (ks. myös Clarke & Moses 2014).

Wallach ja Marchant (2018) ehdottavat **valtakunnallista ja kansainvälistä komiteaa** sidosryhmien välisen yhteistyön, kommunikoinnin ja yhteissääntelyprosessin tukemiseksi. Komitean keskeisenä tarkoituksena olisi tukea ketterän ohjauksen toteutumista. Komitean tulisi edustaa mahdollisimman monia sidosryhmiä teollisuudesta ja kansalaisyhteiskunnasta valtiollisiin toimijoihin ja kansainvälisiin standardointielimiin sekä sellaisia ihmisryhmiä, jotka muutoin olisivat aliedustettuina. Komitealla olisi useita tehtäviä, jotka liittyvät eri sidosryhmien osallistamiseen, yhteisen keskustelufoorumin tarjoamiseen ja eturistiriitojen sovittelemiseen. Se levittäisi ja arvioisi tietoa sekä analysoisi ja tuottaisi ehdotuksia pehmeän ja kovan ohjauksen kehittämiseksi. Myös de Almeida ym. (2021), Cath ym. (2018) ja Floridi ym. (2018) ovat ehdottaneet vastaavanlaisia organisaatiomuotoja, jotka kokoaisivat sidosryhmät yhteen toimien koordinaattorina. Nämä tutkijat ehdottavat edellä mainitun lisäksi, että koordinoivan organisaation tulisi tukea tietojen keräämistä ja analysointia, sen tulisi avustaa ja neuvoa eri sidosryhmiä sosiaalisesti ja ympäristöllisesti kestävän tekoälyn kehittämisessä, tehdä ennakointianalyysejä visioidun ja toivotun tulevaisuuden määrittelemiseksi sekä antaa suosituksia ja toimintaohjeita tekoälyn ohjaukseen liittyen.

3.3 Eettisten ja ihmisoikeusperiaatteiden soveltaminen ohjauksessa

Kolmas aihe käsittelee normatiivisten periaatteiden hyödyntämistä tekoälyn ohjauksessa, jota kuvailemme tässä eettiseksi ohjaukseksi. Normatiiviset periaatteet perustuvat kirjallisuudessa pääosin etiikkaan ja ihmisoikeuksiin, mutta myös hyvään hallintoon ja vastuulliseen tutkimukseen ja innovointiin (RRI). Eettisen ohjauksen haasteena on niiden heikko täytäntöönpano ja omaksuminen tekoälyä koskevassa päätöksenteossa ja organisaatioiden käytännöissä. Tarkastelemassamme kirjallisuudessa haasteeseen esitetään kahdenlaista lähestymistapaa. Ensimmäinen liittyy eettisten periaatteiden toimeenpanon keinoihin (Stix 2021; de Almeida ym. 2021; Tsamados ym. 2021; Delacroix & Wagner 2021; Wallach & Marchant 2018; Floridi ym. 2018; Rahwan 2018). Toinen esittää eettisten periaatteiden sijaan ihmisoikeuksien soveltamiseen perustuvia ratkaisuja (Donahoe & Metzger 2019).

Jälkimmäisen lähestymistavan puoltajien (Donahoe & Metzger 2019; ks. myös Yeung ym. 2019; Smuha 2020) mukaan kansainvälisiin ihmisoikeusnormeihin perustuva ohjaus ja koordinaatio on paras lähtökohta yhteistä hyvää tukevien tekoälyjärjestelmien suunnittelulle, kehittämiselle ja käytölle. Kolme pääasiallista argumenttia puoltaa tätä näkemystä. Ensinnäkin yritysten ja eri organisaatioiden laatimat eettiset ohjeet voivat olla ala- tai organisaatiokohtaisia, eikä niitä ole suunniteltu julkista ohjausta ja päätöksentekoa varten. Vaarana on arvorelativismi. Toiseksi ihmisoikeudet tarjoavat vakiintuneen ja yleismaailmallisen arvopohjan tekoälyn eettiselle kehittämiselle. Ihmisoikeuksia käytetään laajasti nykyhallintojen arvopohjana. Kysymys kuuluukin, miten ihmisoikeuksiin perustuva arvopohja peilautuu käytännön toimissa ja miten se vaikuttaa tekoälyn julkiseen ohjaukseen. Kolmanneksi ihmisoikeuksien yleismaailmallinen julistus soveltuu myös tekoälyn vaikutusten huomioimiseen. Ihmisoikeusperustainen ohjaus voisi siis vastata moniin tekoälyn ohjauksen haasteisiin niin sääntelyn kuin hallinnon suunnittelun ja täytäntöönpanon tasolla.

On huomioitava, että eettisten periaatteiden ja ihmisoikeuksien toimintaedellytykset ovat päällekkäisiä ja täydentävät toisiaan (Stix 2021, 15). Ihmisoikeusperusteista lähestymistapaa puoltavat kirjoittajat eivät sivuuta eettisiä periaatteita, vaan argumentoivat sen puolesta, että periaatteiden tulisi perustua ihmisarvoihin sen sijaan, että ne olisivat koosteita eri yritysten käytännesäännöistä ja kansainvälisten organisaatioiden laatimista eettisistä suosituksista. Taulukossa 4 on tämän vuoksi koottu yhteen eettisten ja ihmisoikeusperiaatteiden käytännön soveltamista tukevat keinot (Taulukko 4).

Normatiivisten periaatteiden toimeenpanoa tukevat keinot voidaan jakaa neljään luokkaan: (1) arviointi, (2) sidosryhmien osallistaminen ja kuuleminen, (3) periaatteiden operationalisoinnin mekanismit (Stix 2021), sekä (4) periaatteiden toimeenpanon valvonta (Sun & Medaglia 2019; Tsamados ym. 2021; de Almeida 2021; Floridi ym. 2018).

Arviointi tukee periaatteiden kehittämistä ja niiden soveltamista erilaisiin asiayhteyksiin sopiviksi. Arviointi auttaa muodostamaan tilannekuvaa

tekoälyjärjestelmän tai palvelun nykytilasta ja tulevaisuuden potentiaalista, eri toimijoiden intresseistä ja näkökulmista teknologiaan sekä sen eettisistä ja ihmisoikeudellisista riskeistä. Arvioimalla nykyisten ohjauskeinojen ja institutionaalisten rakenteiden kyvykkyyttä ehkäistä tekoälyn riskejä ja tukea eettisten periaatteiden toteutumista on mahdollista muodostaa käsitys sääntelyn ja ohjauksen kehittämisen tarpeista. Sidosryhmien osallistuminen ja julkinen keskustelu ovat keskeisiä tekijöitä myös arvioitaessa paikallista teknistä, organisatorista, lainsäädännöllistä ja institutionaalista ympäristöä, jossa tekoälyjärjestelmät toimivat ja joiden puitteissa niiden käyttöä ohjataan. (Stix 2021; de Almeida ym. 2021; Floridi ym. 2018).

Taulukko 4. Eettisten ja ihmisoikeusperiaatteiden käytännön soveltamista tukevat keinot

Tehtävä	Kuvaus	Viitteet
Arviointi	• Arviointi tukee periaatteiden kehittämistä ja niiden soveltamista erilaisiin asiayhteyksiin sopiviksi • Arvioi tekninen kehitys ja siihen liittyvät eettiset riskit sekä erilaisten toimijoiden intressit • Arvioi nykyisten ohjauskeinojen ja institutionaalisten rakenteiden kyvykkyyttä ehkäistä tekoälyn riskejä ja tukea eettisten periaatteiden toteutumista • Sidosryhmien osallistuminen ja julkinen keskustelu ovat arvioinnissa keskeisiä	Stix 2021; Floridi ym. 2018; de Almeida ym. 2021
Yhteistyö ja sidosryhmien osallisuus	• Sidosryhmien laaja yhteistyö ja osallistuminen on keskeinen keino eettisten periaatteiden kehittämisessä ja soveltamisessa. • Osallista sidosryhmiä kaikkiin tekoälyn normatiivisen ohjauksen vaiheisiin: suunnittelu-, toteutus- ja toteutuksen jälkeisiin vaiheisiin. • Osallista ammattijärjestöjä periaatteiden kehittämiseen. • Osallisuus voi olla koordinoivan tahon vastuulla	Stix 2021; Delacroix & Wagner 2021; Donahoe & Metzger 2019; Nemiz 2018

Periaatteiden operationalisointia tukevat mekanismit	• Mekanismit, joilla mahdollistetaan kansalaiskeskustelu ja annetaan kansalaisyhteiskunnalle mahdollisuus vaikuttaa tekoälyä koskeviin päätöksiin. • Ohjausta ja koordinointia tukevien mekanismien ja instituutioiden kehittäminen. • Tekniset ratkaisut ja ei-tekniset suositukset ja suuntaviivat eettisten riskien välttämiseksi tai eettisen pohdinnan edistämiseksi. • Taloudelliset kannustimet eettisten periaatteiden toteutumista tukeville sovelluksille ja menettelyille. • Tekoälyn kehittäjien itsesääntelyn ja eettisten valmiuksien tukeminen. • Tekoälyn ohjauksen ja etiikan kehittäjien tukeminen tekoälyn vaikutusten ymmärtämisessä. • Käytännön soveltamisen edellytyksiä ovat toimivat valvontarakenteet, vastuuvelvollisuus, jäljitettävyys, sanktiomekanismit ja suunnittelu sidosryhmien osallistumisen ja arvojen yhdenmukaistamisen tukemiseksi tekoälyn kehittämisessä ja käytössä.	Stix 2021; Rahwan 2018; Yeung ym. 2021; Wallach & Marchant 2018; Floridi ym. 2018; Tsamados ym. 2021; Sun & Medaglia 2019; de Almeida ym. 2021; Truby 2020

Sidosryhmien osallistuminen on tärkeää tekoälyn normatiivisten periaatteiden määrittelemisessä, niiden toimeenpanossa ja toteutumisen seurannassa (Stix 2021; Delacroix & Wagner 2021). Sidosryhmien ja kansalaisten varhainen osallistaminen ja monialainen asiantuntijavuoropuhelu auttavat varmistamaan, että tekoälyä ohjaavat periaatteet ovat demokraattisesti ja ihmisoikeuksien kannalta oikeutettuja (Donahoe & Metzger 2019; Stix 2021; Nemiz 2018). Eettisten toimintaperiaatteiden pitäisi olla moniäänisen sidosryhmäyhteistyön tulos. Tekoälyn käyttöä ja kehittämistä koskevia eettisiä periaatteita on kyseenalaistettava, jos niitä laatimassa on yksipuolinen joukko toimijoita, kuten yksityinen sektori (Delacroix & Wagner, 2021). Tämä näkökulma pätee myös julkishallintoon tai muihin organisaatioihin, jotka pyrkivät laatimaan tekoälyä koskevia politiikkatoimenpiteitä ilman laajempaa sidosryhmien kuulemista: vaarana on tavoitteiden ja toimenpiteiden yksipuolistuminen tai tietyn eturyhmän vaikutusvallan korostuminen normatiivisten periaatteiden kehystämisessä (Sun & Medaglia 2019). Delacroix ja Wagner (2021) ehdottavat, että julkishallinnon olisi kehotettava myös ammattijärjestöjä osallistumaan normatiivisten periaatteiden ja niiden toimeenpanomekanismien laatimiseen. Erillinen koordinoiva taho voi tukea sidosryhmien osallistumista (Wallach & Marchant 2018; de Almeida ym. 2021; Floridi ym. 2018).

Periaatteiden operationalisointia tukevat työkalut ovat käytännönläheisiä ja konkreettisia ohjeistuksia, kuten työkalupakkeja, menetelmiä tai teknisiä ratkaisuja, joka auttavat periaatteiden ja suositusten käyttöönottamisessa (Stix 2021, 12). Ohjeisiin voi kuulua esimerkiksi menetelmiä ja mekanismeja kansalaiskeskustelun synnyttämiseksi ja sen varmistamiseksi, että kansalaisyhteiskunnalla on mahdollisuus vaikuttaa tekoälyä koskeviin päätöksiin (Stix 2021; Rahwan 2018). Floridi (2018) kumppaneineen ovat esittäneet 20 konkreettista toimenpidettä, joiden

avulla poliittiset päättäjät voivat tukea eettistä tekoälyä. Ne sisältävät nykyisten säännösten ja institutionaalisten valmiuksien **arvioinnin**; oikeudellisten ja koordinoinnin menettelyiden ja näitä tukevien mekanismien ja instituutioiden **kehittämisen**; taloudelliset **kannustimet**, joilla tuetaan sellaisia ohjauskeinoja, toimenpiteitä ja teknisiä sovelluksia, jotka ovat linjassa sosiaalisesti suotavien tavoitteiden kanssa; sekä eettisen valveutuneisuuden, osaamisen ja itsesääntelyn **tukeminen**. Esimerkiksi päättäjien, opiskelijoiden ja ammattilaisten kouluttaminen tietotekniikan, ihmisoikeuksien ja etiikan aloilla nostaa eettistä valveutuneisuutta. Tarvitaan myös monialaista vuoropuhelua sen varmistamiseksi, että päätöksentekijät, tekoälyn kehittäjät, yritykset ja kansalaiset olisivat tietoisia tekoälyjärjestelmien yhteiskunnallisista ja eettisistä vaikutuksista sekä vastuullista toimintaa tukevista konkreettisista suosituksista (Floridi ym. 2018; Donahoe & Metzger 2019; Truby 2020).

Periaatteiden toimeenpanon valvonta on operationalisoinnin edellytys. Viranomaisen tulisi varmistaa tekoälytoiminnan asianmukainen valvonta ja sanktiomekanismit, toiminnan jäljitettävyys, vastuuvelvollisuuden toteutuminen, sekä kanaslaiskeskusteluun ja -vaikuttamiseen perustuva suunnittelu (Wallach & Marchant 2018; Tsamados ym. 2021; Truby 2020). Julkisen hallinnon olisi määriteltävä tarkasti, millaisia tekoälyn ohjaukseen ja valvomiseen liittyviä tehtäviä eri hallintoelinten tulisi ottaa vastuulleen (de Almeida ym. 2021). Jätämme tässä raportissa käsittelemättä ehdotukset teknistä ratkaisuista ja sovelluksia, joiden avulla ohjelmistosuunnittelijat voivat välttää eettisiä haasteita ja tukea eettisten periaatteiden toteutumista (Tsamados ym. 2021; Yeung ym. 2019; Truby 2020).

3.4 Viranomaisen tehtävät tekoälyn ohjauksessa

Eettisten periaatteiden noudattamisen varmistamiseksi tarvitaan valvontaa (esim. Tsamados ym. 2021; Floridi ym. 2018; Wallach & Marchant 2018), esimerkiksi riippumattoman instituution tai viranomaisen toteuttamana (de Almeida ym. 2021; Dignam 2020; Bannister & Connolly 2020, ks. myös Yeung ym. 2019). Toimintaperiaate voisi vastata lääketieteellisellä alalla toimivien lupa- ja valvontaviranomaisten toimintaa, eli se arvioisi, valvoisi ja hyväksyisi algoritmien käytön tietyillä aloilla (Floridi ym. 2018; Dignam 2020; Bannister & Connolly 2020).

Tähän ei välttämättä tarvita erillistä laitosta, vaan valvonnasta voisi vastata eri virastoista tai ministeriöistä koostuva ryhmä (Clarke 2019; de Almeida ym. 2021). Wallach ja Marchantin (2018) näkemyksen mukaan julkisten viranomaisten tulisi niin ikään edistää pehmeiden ohjauskeinojen toimeenpanoa esimerkiksi vaatimalla teollisuuden toimijoita noudattamaan erinäisiä standardeja, kuten laadunhallintaa koskevaa ISO 9000 -standardia. Seuraavassa taulukossa (Taulukko 5) esitetään yhteenveto kirjoittajien ehdottamista tehtävistä viranomaisille.

Taulukko 5. Ehdotuksia tekoälyviranomaisen mahdollisista tehtävistä

Tehtävät	Kuvaus	Viitteet
Algoritmien valvonta ja lupien myöntäminen	Tekoälytuotteiden, -ohjelmistojen, järjestelmien tai palveluiden arviointiin ja valvontaan kuuluu: • suunnittelun, verifioinnin, testauksen ja arvioinnin valvonta ja niitä koskevien vaatimusten toimeenpano • varmistaa, että algoritmit noudattavat ajanmukaisia standardeja, toimivat asianmukaisesti, niitä on testattu, ja niillä on selkeä vastuuvelvollisuusmekanismi	de Almeida ym. 2021; Floridi ym. 2018; Bannister & Connolly 2020
Tekoälyä kehittävien organisaatioiden valvonta	• Valvontaan kuuluu vaatimus yleisen edun mukaisuuden toteutumiseksi organisaatioiden päätöksenteossa • Tekoälyviranomaisella on oikeus vaikuttaa tiettyjen yritysten hallitukseen käyttäen veto-oikeutta esim. immateriaalioikeuksiin ja hallintoneuvoston henkilöstöön liittyvissä asioissa	Dignam 2020
Arviointi	• Viranomaisen tehtäviin kuuluu arvioida tekoälytuotteiden ja niiden kehittämisen eettisiä ja sosiaalisia vaikutuksia, tietohallintoa, riskienhallintamekanismeja ja lainsäädännön vaikutusta toimintaan	de Almeida ym. 2021
Sertifiointi, auditointi ja kehitystyö	• Tuotteiden ja palveluiden sertifiointi ja sertifikaattien hallinnointi sekä näiden osalta yhteydenpito toimeenpanovaltaan ja uuden lainsäädännön tarpeen arviointi • Varmentaa standardien noudattaminen sekä dokumentointi-, avoimuus-, koulutus-, vastuu- ja testausvaatimukset. • Eettisten vaikutusten arviointimenettelyjen, tiedonhallintamallien, mahdollisten vääristyneiden järjestelmien ja riskinhallintamekanismien auditointi. • Tietojen havaitsemisjärjestelmien kehittäminen, riskinhallinta sekä tekoälyn T&K-toiminnan standardointi, sertifiointi ja auditointi. • Vuoropuhelu alan toimijoiden kanssa parhaista käytännöistä ja riskinhallintastandardeista. • Eettisten ongelmien määritelmien ja eettisten vaikutustenarviointien kehittäminen. • Lainsäädännön, politiikan ja teknologian välisen vuorovaikutuksen vahvistaminen	de Almeida ym. 2021; Bannister & Connolly 2020
Testaus ja lisensointi	• Viranomainen varmistaa, että suunnittelussa ja testauksessa ei esiinny vääristymiä. • Lisenssi perustuu arviointiin tekoälyn käyttämisen tarkoituksenmukaisuudesta eri asiayhteyksissä. • Lisenssin lupamaksu määrittyy sen mukaan, minkälaisia sosiaalisia ja yhteiskunnallisia (kuten työllisyysvaikutukset) vaikutuksia tekoälysovelluksella on	Dignam 2020

4 Kohti kokonaisvaltaista, osallistavaa, institutionalisoitua ja soveltuvaa tekoälyn julkista ohjausta ja koordinaatiota

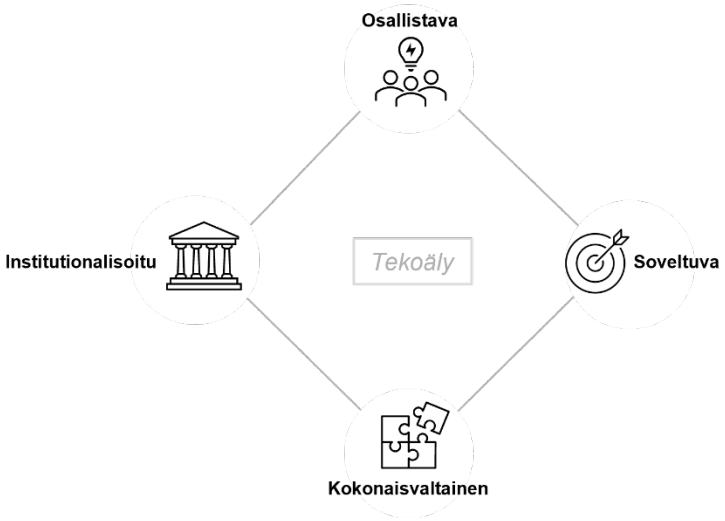
Katsauksessamme kootut ehdotukset tarjoavat julkishallinnolle keinovalikoiman yhteistä hyvää tuottavan tekoälyn tukemiseksi sekä keskeiset tavoitteet tekoälyä koskevalle päätöksenteolle. Ehdotukset koostuvat tutkimuskirjallisuudessa esiintyvistä keinoista ja menettelyistä, joita noudattamalla pyritään varmistamaan, että toiminta on vastuullista ja perustuu eettiseen harkintaan. Koottuna ehdotukset edustavat useiden tutkijoiden näkökulmaa tavoittelemisen arvoisista eettisen ohjauksen toteuttamisen edellytyksistä ja tavoista liikkua yleisluontoisista ja abstrakteista tekoälyn eettisistä periaatteista kohti niiden toimeenpanoa tukevia konkreettisia menettelyitä paikallisissa asiayhteyksissä. Tulosten yhteenvetona esitämme neljä ristikkäistä ideaalia tavoitetta tekoälyn julkiselle ohjaukselle ja koordinaatiolle (kuva 3):

(1) ***Kokonaisvaltainen*** – Tarvitaan tekoälyilmiön kokonaisvaltaisuuden ja systeemisen luonteen huomioiva hallintamalli, joka toimii vertailupisteenä julkisille hallintoelimille ja muille asiaankuuluville toimijoille ja helpottaa tekoälyn vastuullisuuden ja etiikan pohtimista paikallisissa asiayhteyksissä kehittäjien ja käyttäjien kanssa.

(2) ***Osallistava*** – Eri sidosryhmien näkemysten ja arvojen huomioiminen vuorovaikutteisessa prosessissa on mukautuvien ja sosiaalisesti hyväksyttävien ohjauskeinojen ja eettisten periaatteiden suunnittelun ja toimeenpanon lähtökohta.

(3) ***Institutionalisoitu*** – Tarvitaan vakiintunutta viranomaistahoa koordinoimaan päätöksentekoa, kehittämään ohjauskeinoja, valvomaan ja varmistamaan lainsäädännön noudattamista ja tukemaan eettisten normien toteutumista.

(4) ***Soveltuva*** – Tarvitaan konkreettisia käytännön ohjeita ja keinoja eettisen periaatteiden ja ihmisoikeuksien soveltamiseksi tekoälyä koskevassa päätöksenteossa sekä paikallisissa asiayhteyksissä.



Kuva 3. Neljä ideaalia tekoälyn julkiselle ohjaukselle ja koordinaatiolle

Osallisuus. Kompleksisten ja nopeasti muuttuvien systeemien ohjaus edellyttää joustavia, paikallisiin asiayhteyksiin soveltuvia ja keskeisten toimijoiden kanssa neuvoteltuja ratkaisuja (Wallach & Marchant 2018; Winfield & Jirotka 2018; Sun & Medaglia 2019; Liu & Maas 2021; Ulnicane ym. 2021). Parhaiden ratkaisujen hakemista edistetään luottamuksen, yhteisymmärryksen ja konsensuksen rakentamisella sidosryhmien kesken sekä dialogisesti pohtimalla teknologian mahdollisia eturistiriitoja, eri arvojen toteutumista ja teknologian vaikutuksia. Tiedon epäsymmetriaa ja epävarmuutta poliittisten päätösten valmistelussa voidaan vähentää huomioimalla eri sidosryhmien tarpeita ja etuja sekä niiden erityistietämystä paikallisista sosiaalisista ja teknisistä ilmiöistä. Tavanomaista laajemmalla sidosryhmänäkemyksellä voidaan tasapainottaa yksityisten yritysten vaikutusvaltaa tekoälypolitiikan, tekoälystrategioiden ja eettisten linjausten kehystämisessä. Osallisuus ja sen edellyttämä järjestäytyminen voisi olla yksittäisen organisaation tai elimen vastuulla, joka koordinoi toimintaa, kouluttaa sidosryhmiä, kerää tarpeellisen tiedon ja tuo eri sidosryhmät keskinäiseen vuoropuheluun.

Soveltuvuus. Tekoälyn etiikan vaikuttavuuden yhtenä ongelmana on, etteivät eettiset periaatteet riittävästi ohjaa poliittista päätöksentekoa tai käytännön toimintaa. Tämä puolestaan murentaa pohjaa luottamukselta, joka on edellytys tekoälyratkaisujen laajamittaiselle käytölle (AI HLEG 2019). Tarvitaan työkaluja ja ohjeita helpottamaan eettisen harkinnan ja periaatteiden soveltamista erilaisissa asiayhteyksissä. Käytännössä työkalut voivat olla toimintaohjeita, ratkaisukortteja, työkaluja ja ohjelmistoja, tai muita menetelmiä ja kannustimia, jotka helpottavat toimintaohjeiden tai periaatteiden soveltamista. Kirjallisuus tarjoaa laajan kirjon ohjeita esimerkiksi kansalaisosallisuuden tukemiseen, arviointityökalujen suunnittelemiseen ja taloudellisten kannustimien luomiseen (Floridi ym. 2018;

Rahwan 2018; Wallach & Marchant 2018; Donahoe & Metzger 2019; Sun & Medaglia 2019; de Almeida ym. 2021; Delacroix & Wagner 2021; Stix 2021; Tsamados ym. 2021). Pelkät työkalut eivät sellaisenaan riitä, vaan tarvitaan myös keinoja työkalujen jalkauttamiseksi osaksi päätöksentekoa. Työkalujen soveltuvuutta ja mielekkyyttä voidaan tukea arvioimalla tekoälyjärjestelmän tai palvelun nykytilaa ja tulevaisuuden vaikutuksia, eri toimijoiden intressejä ja näkökulmia, järjestelmän eettisiä ja ihmisoikeudellisia riskejä sekä nykyisten ohjauskeinojen ja rakenteiden kyvykkyyttä ehkäistä tekoälyn riskejä ja tukea eettisten periaatteiden toteutumista. Toiseksi tarvitaan sidosryhmien osallistamista eettisen periaatteiden jalkauttamisen ohjeiden suunnitteluun, käytännön kokeilemiseen ja toteutukseen sekä toteutuksen jälkeiseen vaiheeseen, jonka myötä voidaan toimintaa kehittää. Kolmanneksi jalkauttamista tukevat menetelmät ja työkalut tulisi integroitua osaksi organisaation tai hallinnon osaston vastuullisuus- tai vastuuvelvollisuuden toimintaohjelmaa, jonka toteutumista sellaisenaan ohjataan erilaisin valvonta-, täytäntöönpano- ja seuraamusmekanismeilla.

Kokonaisvaltaisuus. Tekoälyn hallintamallin tulisi auttaa viranomaisia ja organisaatioita huomioimaan tekoälyilmiön systeeminen ja moniulotteinen luonne paikallisessa päätöksenteossa (Wirtz ym. 2020; Gasser & Almeida 2018; de Almeida ym. 2021; Yeung ym. 2019). Hallintamallin olisi tuettava kokonaisvaltaisen ja integroidun ohjauskeinovalikoiman kehittämistä, jossa tekoälyn ohjauksen eri ulottuvuudet lainsäädännöstä eettisiin periaatteisiin ja teknisiin ratkaisuihin nähdään systeemisenä kokonaisuutena. Koska tekoäly on geneerinen teknologia, jota voidaan soveltaa hyvin erilaisissa asiayhteyksissä ja teknologiakokonaisuuksissa, on tärkeää, että malli tekee paikallisen monimuotoisuuden ja sen edellyttämän ohjauksen erityiskysymykset näkyväksi. Tekoäly on siis ymmärrettävä kontekstisidonnaisena ilmiönä, joka edellyttää ohjauskeinoilta ketteryyttä. Esimerkiksi missiolähtöinen innovaatiopolitiikka tarjoaa yhden lähestymistavan toimintatapojen uusimiselle kohti systeemisempää ja ketterämpää ohjausta (Lähteenmäki-Smith 2020).

Institutionalisointi. Tekoälyilmiön kokonaisvaltaisen julkisen ohjauksen koordinoimiseksi, päällekkäisyyksien välttämiseksi ja ad hoc politiikkatoimenpiteiden välttämiseksi tekoälyn ohjaus tulisi kokonaisuutena liittää olemassa olevan hallinnon yksittäisen osaston tai osastojen vastuulle, tai tehtävää varten tulisi perustaa erillinen virasto (Floridi ym. 2018; Bannister & Connolly 2020; Dignam 2020; de Almeida ym. 2021). Tällaisen viraston tehtävänä olisi päätöksenteon koordinointi sekä soveltuvien ja toteuttamiskelpoisten ohjauskeinojen jatkuva kehittäminen. Virasto toimisi eräänlaisena lupa- ja valvontaviranomaisena, joka valvoisi ja varmistaisi lainsäädännön noudattamista, tekoälyjärjestelmien turvallisuutta ja normatiivisten periaatteiden toteutumista. Sen vastuualueelle voisi kuulua tekoälyjärjestelmien asianmukaisuuden valvonta, verifiointi, sertifiointi, auditointi, riskien hallinta, sekä lainsäädännön kehittäminen. Lisäksi virasto voisi toteuttaa vaikutustenarviointia, jota tarvitaan lainsäädännön kehittämiseksi sekä tekoälyn etiikkaa koskevan julkisen keskustelun ylläpitämiseksi.

5 Lopuksi

Tässä raportissa esitettyjen kehitysehdotusten perusteella voimme konkretisoida ajatusta julkishallinnon roolista eettisen tekoäly-yhteiskunnan rakentamisessa. Kuten olemme johdannossa maininneet, tekoälyn eettistä ohjausta ei ole tarkkaan määritelty tekoälyn ohjausta koskevassa kirjallisuudessa, vaan sillä viitataan laajasti ohjaus- ja koordinaatiokeinoihin, joiden tavoitteena on minimoida tekoälyn riskejä ja tukea teknologian käyttöä yhteisen hyvän sekä sosiaalisen, taloudellisen ja ekologisen kestävyyden hyväksi (Ireni-Saban & Sherman 2021; Mazzi & Floridi 2023; Stahl 2021; Taddeo & Floridi 2018; Winfield & Jirotka 2018). Raportissa esitettyjen kehitysehdotusten ja analysoidun kirjallisuuden perusteella voimme tiivistäen todeta, että tekoälyn julkisen eettisen ohjauksen tulisi hyödyntää etiikkaan, ihmisoikeuksiin, hyvään hallintoon sekä vastuulliseen tutkimukseen ja innovointiin perustuvia normatiivisia periaatteita politiikanteossa. Ohjauksen olisi perustuttava osallistaviin ja joustaviin hallintomekanismeihin ja -käytäntöihin, joiden tarkoituksena on tukea päätöksentekoa ja sopeutumiskykyä lyhyen ja pitkän aikavälin tekoälyn yhteiskunnallisiin haasteisiin. Julkisessa päätöksenteon valmistelussa olisi tiedostettava ohjauskeinojen kehystäminen ja olla refleksiivisiä sen suhteen, kuka osallistuu ja pääsee vaikuttamaan tekoälyä koskeviin strategioihin ja päätöksiin. Tekoälyn ohjauksen sosiaalisesti kestävä ja pitkäaikainen kehittäminen edellyttää institutionalisoitua koordinoitua ja päätöksenteon rakennetta, joka tuottaa tietoa tekoälyn vaikutuksista ja sidosryhmänäkemyksistä, koordinoi tekoälyn ohjausta, valvoo normatiivisten periaatteiden toteutumista, tekoälysovellusten turvallisuutta ja riskienhallintaa. Konkreettisimmillaan näitä tavoitteita tukevia käytäntöjä on kuvattu edellä esitetyissä taulukoissa 1–5.

Ehdotetut ideaalit tarjoavat suuntalinjoja kokonaisvaltaiseen mutta joustavasti käyttötilanneittain soveltuvaan päätöksentekoon. Ideaalien taustaoletuksena on, että ohjauksessa sovellettujen menettelyiden ja sääntöjen arvoperusta ja siinä hyödynnettävän informaation laatu määrittää pitkälti eettisen ja vastuullisen toiminnan mahdollisuuksia. Eettisen ja vastuullisen toiminnan arvoperustana tulisi olla hyvän hallinnon ja ohjauksen periaatteet, ihmisoikeudet sekä vastuullisen tutkimus ja innovaatiotoiminnan periaatteet. Päätöksenteossa olisi hyvän tietopohjan varmistamiseksi huomioitava laajasti eri sidosryhmien näkemyksiä sekä ennakoinnin ja vaikutustenarvioinnin kautta saatavaa tietoa päätösten lyhyistä ja pitkän aikavälin tavoitteista. Näiden perusteella tulisi laatia pitkän aikavälin

strategia, joka ohjaa julkista päätöksentekoa suhteessa vastuullisuus- ja kestävyystavoitteisiin. Ohjeita noudattamalla voidaan tukea kestävää ja luottamusta synnyttävää tekoälypolitiikkaa ja tekoälypalveluita.

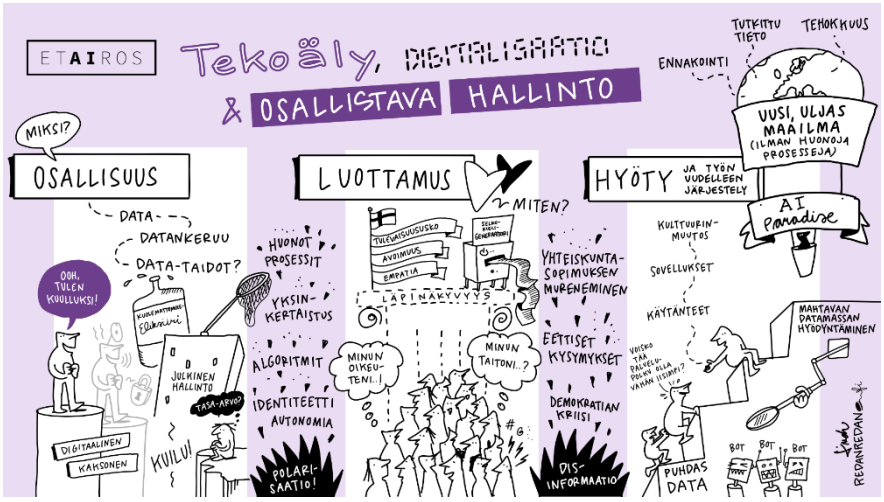
Julkisen hallinnon tehtävänä on huomioida erilaisia yhteiskunnallisia tarpeita ja arvoja ja noudattaa suhteellisuusperiaatetta ja kohtuullisuutta erilaisten intressien ja yhteiskunnallisten arvojen tasapainottamisessa. Samalla on otettava huomioon tekoälysovellusten eettiset, lainsäädännölliset ja yhteiskunnalliset vaikutukset. Kuten olemme esittäneet, tämä edellyttää ketteriä ja mukautuvia ohjauskeinoja, joiden myötä tekoälyn ohjauksen käyttötilanteet punoutuvat eri sidosryhmien jaettuun näkemykseen sekä keskusteluun tekoälyn yhteiskunnallisista vaikutuksista.

Kehitysehdotukset ovat tiiviisti yhteydessä laajempaan keskusteluun valtion roolista teknologian ja TKI politiikan ohjaajana ja toiminnan paradigmaattisista, poliittisista ja ideologisista lähtökohdista. Parhaimmillaan julkisessa ohjauksessa hyödynnetään laajasti käsillä olevaa eri hallintaparadigmoista koostuvaa keinovalikoimaa, yhdistäen ja soveltaen horisontaalista koordinaatiota eri organisaatioiden ja sidosryhmien osallistamiseksi sekä vertikaalista valvontaa aineellisten ja proseduraalisten normien täytäntöönpanon varmistamiseksi. Eettisten ja vastuullisuusperiaatteiden käytäntöön vieminen edellyttää julkishallinnolta hierarkkisen päätöksenteon sijaan osallisuutta poliittisessa päätöksenteossa. Sidosryhmien ja kansalaisten osallisuus päätöksentekoprosessissa oikein hyödynnettynä vähentää informaation asymmetrioita, sekä parantaa päätöksenteon legitimiteettiä ja ketteryyttä (Delacroix & Wagner 2021; Stix 2021). Viranomaisen olisi kuitenkin valvottava sovittujen normien toimeenpanoa, jotta erilaisten turvallisuuteen, yksityisyyden suojaan, luotettavuuteen, etiikkaan ja vastuullisuuteen liittyvien normien toteutuminen voidaan varmistaa.

Tekoälyn julkisen ohjauksen eteen on tehtävä vielä paljon työtä. Raportissa esitetyt ohjauksen kehittämisehdotukset eivät perustu empiirisiin tutkimuksiin niiden toimeenpanon vaikuttavuudesta, minkä vuoksi olisikin tärkeä tutkia, millä tavalla normatiiviset periaatteet käytännössä ohjaavat tekoälyn käyttöä ja kehitystä yhteisen hyvän puolesta, ja miten ohjausta ja politiikkatoimenpiteitä voisi kehittää empiiriseen tietoon ja kokeiluihin perustuen. Kokeilut ja tutkimukset ohjauksen käytännöistä erilaisissa teknologisissa, hallinnollisissa ja sosiaalisissa yhteyksissä saattaisivat merkittävästi rikastaa keskustelua tekoälyn julkisen ohjauksen kehittämisestä.

5.1 Miten osallistavaa hallintoa voisi kehittää tekoälyn avulla? Näkökulmia ja nostoja julkishallinnon sparrausklinikasta

Tekoäly ja digitaaliset kansalaisosallisuuden teknologiat tarjoavat uusia näkymiä päätöksenteon kehittämiseen. Julkishallinnon näkökulmasta niiden hyödyntäminen edellyttää tasapainoilua. Tasapainoilun täytyy olla kuitenkin luonteeltaan harkitsevaa, kokeilevaa ja rohkeaa.



Linda Saukko-Raudan tulkinta sparrausklinikan keskusteluosion annista. ©ETAIROS.

ETAIROS tutkimushanke järjesti toukokuussa 2023 strategisen tutkimuksen neuvoston rahoittaman STEER-ohjelman puitteissa julkishallinnon sparrausklinikan tekoälystä, digitalisaatiosta, ja osallistavasta hallinnosta (ks. yhteenveto tapahtumasta). Työpajassa nousi esille, että kansalaiset toivovat digitaalisia vaikuttamisen keinoja, kun taas hallinto toivoo kasvokkain tapahtuvaa vuorovaikutusta (Jämsén ym. 2022). Syitä tähän on helppo spekuloida. Osallistuminen digitaalisilla alustoilla voi olla helppoa ja aika- ja paikkariippumatonta. Toisaalta nykyiset digitaaliset alustat rajaavat kommunikaatiomuotona pois merkittäviä dialogisuuden onnistumisen ehtoja, kuten esimerkiksi tunnetilojen ja nyanssin välittämistä. Tämä ei tietenkään tarkoita sitä, etteikö teknologiaa tai sen käyttötapoja voisi tässä suhteessa kehittää.

Kysyntä ”demokratiateknologialle” on joka tapauksessa nousussa. Tuore raportti arvioi osallistamiseen, harkintaan ja äänestämiseen liittyvän teknologian markkinoiden kasvavan Euroopassa 800 miljoonaan (€) seuraavan viiden vuoden aikana (García ym. 2023). Erään viitekehyksen keskustelulle tarjoaa myös OECD:n raportti vuodelta 2021. Raportin mukaan Suomessa sekä julkinen sektori että

kansalaiset ovat passiivisia toistensa suhteen, ja ”Suomen paradoksi” on tästä huolimatta kansalaisten parissa vallitseva korkea luottamus hallintoa kohtaan. Osallistamiselle ja sitä mahdollistavalle teknologialle voidaan nähdä kysyntää siis Suomessakin. Esimerkiksi Helsingin kaupunki on jo käyttänyt digitalisaatiota ja tekoälyä hyödyksi nuorten osallistamiseen budjettiratkaisuissa – tavoittaen jopa kolmasosan kohdeyleisöstään (Giesen 2023).

Millainen on tulevaisuuden demokratia tai esimerkiksi älykäs kaupunki? Millaisen haluamme niiden olevan? Ei liene erimielisyyttä siitä, että osallistamista ei kannata tehdä osallistamisen vuoksi. On vielä hahmottumatonta, mitä uusilla kansalaisteknologioilla voitaisiin parhaimmillaan saavuttaa, ja kuka olisi vastuussa asian edistämisestä. Mikä olisi julkisen hallinnon motiivi muuttaa nykyisiä käytänteitä osallistavammiksi? Tekoälyyn liittyvät eettiset ja juridiset reunaehdot luovat perustan sen soveltamiselle. Niiden suhteen alkaa olla myös kiire. Klinikassa arvioitiin, että aikaa on kahdesta kolmeen vuotta saada juridinen kehys kuntoon. Päättäjien ”sense of urgency” täytyy nousta tekoälyn suhteen – tekoälystä on otettava hyöty irti niin kansakunnalle kuin koko Euroopallekin. Toisaalta tekoälyn vastuullinen soveltaminen edellyttää myös rauhallista harkintaa – ja osallistamista. Teemaa ei tule kuitenkaan mystifioida – kyse on myös olemassa olevien keinojen ja jännitteiden (niin teknisten kuin hallinnollisten) nostamisesta pöydälle ja keskustelun aloittamisesta. Kenties asia ei lopulta ole ”sen kummempi”, klinikassa todettiin. Tekoäly voikin toimia tapana nostaa esille keskustelua erilaisten prosessien uudistamisesta ja niiden tarpeesta, sekä tekniikan roolista.

Sparrausklinikassa todettiin, että digitalisaatioon liittyy hyviä aikeita, mutta niiden todellinen arvo saattaa piillä välineenä prosessien ja toimintatapojen uudistamiseen. Tekoälykään ei ole kaikkivoipaa, vaan edelleen kehittyvää teknologiaa. Esimerkiksi ChatGPT tuottaa aina jonkin vastauksen syötteeseen todennäköisyyslaskennan perusteella, ellei siihen ole erikseen rakennettu kyseisen tematiikan suhteen jotakin säännöstöä, joka estäisi vastineen generoinnin. ChatGPT:n käyttö edellyttää siis taitoja, mutta millaisia – sitä ei olla vielä täysin kyetty jäsentämään. Tekoälyn käyttöä myös säännellään. On muistettava, että esimerkiksi 1. toukokuuta 2023 voimaan astunut hallintolain pykälä 53 e estää automaattisen päätöksenteon sellaisissa tapauksissa, jotka edellyttävät tapauskohtaista harkintaa. Yhdessä nämä tarkoittanevat, että käyttäjien täytyy ymmärtää tekoälyteknologiaa ja sen teknisiä ja juridisia rajoja. Tekoälyn eettinen ja vastuullinen käyttö on lopulta osa tulevaisuuden hyvää hallintotapaa.

Kokeilukulttuurin on alettava nyt.

Kirjallisuus

AI HLEG (2019). High-level expert group on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (accessed January 14, 2022).

Asaduzzaman, M. & Virtanen, P. (2016). “Governance theories and models,” in Global Encyclopedia of Public Administration, Public Policy, and Governance, ed. A. Farazmand. New York, USA: Springer. doi: 10.1007/978-3-319-31816-5_2612-1

Baldwin, R., Cave, M. & Lodge, M. (2012). Understanding regulation: Theory, strategy, and practice (2nd ed.). Oxford University Press. doi:10.1093/acprof:osobl/9780199576081.001.0001

Bannister, F. & Connolly, R. (2020). Administration by algorithm: A risk management framework. Information Polity, 25(4), 471–490. doi: 10.3233/IP-200249

Bareis, J. & Katzenbach, C. (2021). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. Science, Technology, & Human Values. doi: 01622439211030007

Baum, S. (2017) On the promotion of safe and socially beneficial artificial intelligence. AI & Soc 32, 543–551. https://doi.org/10.1007/s00146-016-0677-0

Bietti, E. (2021). "From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics," in Journal of Social Computing, vol. 2, no. 3, pp. 266-283, doi: 10.23919/JSC.2021.0031.

Borrás, S. & Edler, J. (2020). The roles of the state in the governance of socio-technical systems’ transformation. Research Policy 49(5), 103971. doi:10.1016/j.respol.2020.103971

Bostrom N. (2014). Superintelligence: Paths, dangers, strategies. New York: Oxford University Press.

Botes, M. (2022) Autonomy and the social dilemma of online manipulative behavior. AI Ethics. https://doi.org/10.1007/s43681-022-00157-5

Buhmann, A. & Fieseler. C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. Technology in Society 64. doi: 10.1016/j.techsoc.2020.101475

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. (2018). Artificial intelligence and the “good society”: the US, EU, and UK approach. Science and Engineering Ethics 24(2), 505–528. DOI: 10.1016/j.techsoc.2020.101475

Christian, B. (2020). The alignment problem: Machine learning and human values. WW Norton & Company.

Cihon, P. (2019). Standards for AI governance: International standards to enable global coordination in AI research & development. Future of Humanity Institute. University of Oxford.

Clarke, R. (2019). Regulatory alternatives for AI. Computer Law & Security Review, 35(4), 398–409. doi: 10.1016/j.clsr.2019.04.008

Clarke, R., & Moses, L. (2014). The regulation of civilian drones’ impacts on public safety. Computer Law & Security Review, 30(3), 263–285. doi:10.1016/j.clsr.2014.03.007

Coeckelbergh, M. (2020). AI ethics. The MIT press essential knowledge series. Cambridge: MIT Press.

Crawford, K. (2021). The atlas of AI. Yale University Press.

Cronin, M. A., & George, E. (2020). The why and how of the integrative review. Organizational Research Methods. doi: 10.1177/1094428120935507

Dafoe, A. (2018). AI governance: A research agenda. Governance of AI Program. Future of Humanity Institute. University of Oxford. Available at: https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf (accessed December 1, 2021).

de Almeida, P. G. R., dos Santos, C. D. & Farias, J. S. (2021). Artificial intelligence regulation: A framework for governance. Ethics and Information Technology 23, 505–525. doi: /10.1007/s10676-021-09593-z

Delacroix, S. & Wagner, B. (2021). Constructing a mutually supportive interface between ethics and regulation. Computer Law & Security Review 40. doi: 10.1016/j.clsr.2020.105520

Dignam, A. (2020). Artificial intelligence, tech corporate governance and the public interest regulatory response. Cambridge Journal of Regions, Economy and Society 13(1), 37–54. doi: 10.1093/cjres/rsaa002

Djeffal, C., Siewert, M. B. & Wurster, S. (2022). Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies. Journal of European Public Policy, 1-23.

Donahoe, E. & Metzger, M. M. (2019). Artificial intelligence and human rights. Journal of Democracy 30(2), 115–126. DOI: 10.1353/jod.2019.0029

Esposito, E. (2022). Artificial Communication: How Algorithms Produce Social Intelligence. The MIT Press.

European Commission (2021a). A European approach to artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (accessed December 29, 2021).

European Commission (2021b). Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (accessed December 29, 2021).

European Commission. (2020). White Paper on Artificial Intelligence: a European approach to excellence and trust. Brussels. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

Floridi, L. & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review 1(1). doi: 10.1162/99608f92.8cd550d1

Floridi, L. (2018). Soft ethics and the governance of the digital. Philosophy & Technology 31(1), 1–8. doi: 10.1007/s13347-018-0303-9

Floridi, L. (2019) Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philos. Technol. 32, 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines 28(4), 689–707. doi: 10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C. &Taddeo, M. (2020). How to design AI for social good: Seven essential factors. Science and Engineering Ethics 26, 1771–1796. doi:10.1007/s11948-020-00213-5

Frederickson, H. G. (2007). “Whatever happened to public administration? Governance, governance everywhere” in The Oxford Handbook of Public

Management, ed. E. Ferlie, L. E. Lynn Jr., and C. Pollitt (New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199226443.003.0013

Future of Life Institute. (2021). AI policy challenges and recommendations. https://futureoflife.org/ai-policy-challenges-and-recommendations/ (accessed December 29, 2021).

Gahnberg, C. (2021). What rules? Framing the governance of artificial agency. Policy and Society 40(2), 194–210.

García, D., Alberto, F., Giesen, L., Landi, M., Mackisack, D., Neven, M., Pearce-Laanela, T., Snehotta, R., Thomas, L., Schwartz, R., Strasser, E., Stühlinger, N., Van der Staak, S., Wetherall-Grujic, G., & Wolf, P. (2023). Democracy Technologies in Europe. Online Participation, Deliberation and Voting. A report for lawmakers, governments and policymakers at national and European level. The International Institute for Democracy and Electoral Assistance (International IDEA), The Innovation in Politics Institute.

Gasser, U. & Almeida, V. A. (2017). A layered model for AI governance. IEEE Internet Computing 21(6), 58–62. doi:10.1109/MIC.2017.4180835

Giesen, L. (2023). One Third of Eligible Youth Participates in Helsinki Youth Budget. Democracy Technologies. https://democracy-technologies.org/participation/one-third-of-eligible-youth-participates-in-helsinki-youth-budget/. [7.6.2023]

Goertzel, B. (2007). Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's critique of Kurzweil. Artificial Intelligence, 171(18), 1161-1173.

Gutierrez, C. I., Marchant, G. E. & Michael, K. (2021). Effective and trustworthy implementation of AI soft law governance. IEEE Transactions on Technology and Society, 2(4), 168-170.

Gutierrez, C. I., Marchant, G. E. & Michael, K. (2021). Effective and trustworthy implementation of AI soft law governance. IEEE Transactions on Technology and Society 2(4), 168–170. doi: 10.1109/TTS.2021.3121959

Hagendorff, T. (2020). “The ethics of AI ethics: An evaluation of guidelines.” Minds and Machines 30, 99–120. doi: 10.1007/s11023-020-09517-8

Hallamaa, J. & Kalliokoski, T. (2022). AI Ethics as Applied Ethics. Frontiers in Computer Science.

Ireni-Saban, L. & Sherman, M. (2021). Ethical governance of artificial intelligence in the public sector. London: Routledge.

Janssen, M. & van der Voort, H. (2016). Adaptive governance: Towards a stable, accountable and responsive government. Government Information Quarterly, 33(1), 1–5. doi: https://doi.org/10.1016/j.giq.2016.02.003

Jasanoff, S. (2016). The ethics of invention: Technology and the human future. WW Norton & Company.

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence 1, 389–399. doi: 10.1038/s42256-019-0088-2

Jämsén, P., Kaartinen, J., Westinen, J. & Turja, T. (2022). Demokraattiset osallistumismahdolli-suudet Suomessa. Kyselytutkimus kansalaisten ja päättäjien ajatuksista päätöksentekoon osallistumisesta ja demokratian tulevaisuuskuvista. Sitran selvityksiä 220.

Karvonen, A., Leikas, J., & Sigfrids, A. (2023). Portaissa tekoälyparatiisiin on vielä nousemista – näkökulmia osallistavan julkishallinnon sparrausklinikasta. Etairos blogi. https://etairos.fi/2023/05/16/portaissa-tekoalyparatiisiin-on-viela-nousemista-nakokulmia-osallistavan-julkishallinnon-sparrausklinikasta/ [7.6.2023]

Konrad, K. & Böhle, K. (2019). Socio-technical futures and the governance of innovation processes: An introduction to the special issue. Futures 109, 101–107. doi: 10.1016/j.futures.2019.03.003

Koskimies, E., Nieminen, M., Stenvall, J., Hallamaa, J., Leikas, J., & Salo-Pöntinen, H. (2021). Suomeen tarvitaan koordinoitu, vastuullisuutta korostava tekoälypolitiikka. Etairos Policy Brief.

Kuhlmann, S., Stegmaier, P. & Konrad, K. (2019). The tentative governance of emerging science and technology: A conceptual introduction. Research Policy 48(5), 1091–1097. doi: 10.1016/j.respol.2019.01.006

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. Asian Journal of Law and Society 7(3), 437–451. doi: 10.1017/als.2020.19

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Lehoux, P., Miller, F. A. & Williams-Jones, B. (2020). Anticipatory governance and moral imagination: Methodological insights from a scenario-based public deliberation study. Technological Forecasting and Social Change 151, 119800. doi: 10.1016/j.techfore.2019.119800

Liu, H. Y. & Maas, M. M. (2021). “Solving for X?” Towards a problem-finding framework to ground long-term governance strategies for artificial intelligence. Futures 126, 102672. doi: 10.1016/j.futures.2020.102672

Lähteenmäki-Smith, K. & Virtanen, P. (2020). ”Mission-oriented public policy and the new evaluation culture,” in Society as an Interaction Space: a systemic approach, ed. Lehtimäki, H., Uusikylä, P., & Smedlund, A., (eds). Society as an interaction space: A systemic approach translational systems sciences, vol 22. Springer, Singapore. doi: https://doi.org/10.1007/978-981-15-0069-5_4

Lähteenmäki-Smith, K., Samuli, M., Vartiainen, P., Uusikylä, P., Jalonen, H., Kotiranta, S., Lintinen, U., Annala, M., Gronchi, I., Leppänen, J. & Mertsola, S. (2021). Valtion ohjaus 2020—luvulla. Säädös—ja resurssiohjauksesta järjestelmänavigointiin. Valtioneuvoston Selvitys—ja Tutkimustoiminnan Julkaisusarja, 2021, 17.

Mager, A. & Katzenbach, C. (2021). Future imaginaries in the making and governing of digital technology: Multiple, contested, commodified. New Media & Society 23(2), 223–236. doi: 10.1177/1461444820929321

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and ‘the dark side’of AI. European Journal of Information Systems, 31(3), 257-268.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI? Nature Machine Intelligence 1, 501–507. doi: 10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics 26(4), 2141–2168. doi: 10.1007/s11948-019-00165-5

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. AI and Ethics, 2(4), 603-609.

Naudé, W. & Dimitri, N. (2020) The race for an artificial general intelligence: implications for public policy. AI & Soc 35, 367–379. https://doi.org/10.1007/s00146-019-00887-x

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180089. doi: 10.1098/rsta.2018.0089

Nieminen. M, Gotcheva, N., Leikas, J. & Koivisto, R. (2019). Ethical AI for the governance of the Society: Challenges and opportunities. In CEUR Workshop Proceedings (Vol. 2505, pp. 20-26).

Nowell, L. S., Norris, J. M., White, D. E. & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. International Journal of Qualitative Methods. https://doi.org/10.1177/1609406917733847

OECD (2019). OECD AI Principles overview. https://oecd.ai/en/ai-principles [18.4.2023]

OECD (2021), Drivers of Trust in Public Institutions in Finland, OECD Publishing, Paris, https://doi.org/10.1787/52600c9e-en. [7.6.2023]

Okoli, C. (2015). A guide to conducting a standalone systematic literature review. Communications of the Association for Information Systems 37(1), 43. doi: 10.17705/1CAIS.03743

Page, M. J., McKenzie, J. E., Bossuyt P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ym. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. The British Medical Journal 2021;372:n71. doi: 10.1136/bmj.n71

Radu, R. (2021). Steering the governance of artificial intelligence: National strategies in perspective. Policy and Society 40:2, 178–193. doi: 10.1080/14494035.2021.1929728

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology 20(1), 5–14. doi: 10.1007/s10676-017-9430-8

Rhodes, R. A. (2016). Recovering the craft of public administration. Public Administration Review 76(4), 638–647.

Rittel, H. W. J. & Webber, M. M. (1973). "Dilemmas in a general theory of planning". Policy Sciences 4(2), 155–169. doi:10.1007/bf01405730

Russell, S. & Norvig, P. (2009). Artificial intelligence: A modern approach (3rd ed.). New Jersey: Pearson Eductation, Inc.

Selbst, A. D. (2017). Disparate impact in big data policing. Ga. L. Rev., 52, 109.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems 10(4), 1–31. doi: 10.1145/3419764

Sigfrids, A., Nieminen, M., Leikas, J., & Pikkuaho, P. (2022). How should public administrations foster the ethical development and use of artificial intelligence? A review of proposals for developing governance of AI. Front. Hum. Dyn. 4:858108. doi: 10.3389/fhumd.2022.858108.

Silverman, C., Mac, R. & Dixit, P., 2020. Facebook ignore political manipulation. Buzzfeed.
https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo [18.4.2023]

Smuha, N. A. (2020). Beyond a human rights-based approach to AI governance: Promise. Philosophy & Technology 34(1), 1–14. doi: 10.1007/s13347-020-00403-w

Stahl, B. (2021). Artificial intelligence for a better future. An ecosystem perspective on the ethics of AI and emerging digital technologies. Springer Briefs in Research and Innovation Governance. doi: 10.1007/978-3-030-69978-9

Stahl, B. C., Andreou, A., Brey, P., Hatzakis, T., Kirichenko, A., Macnish, K., Laulhé Shaelou, S., Patel, A., Ryan, M. & Wright, D (2021). Artificial intelligence for human flourishing: Beyond principles for machine learning. Journal of Business Research 124, 374–388. doi: 10.1016/j.jbusres.2020.11.030

Stix, C. (2021). Actionable principles for artificial intelligence policy: Three pathways. Science Engineering Ethics 27(1), 1–17. doi: 10.1007/s11948-020-00277-3

Sun, T. Q. & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly 36(2), 368–383. doi: 10.1016/j.giq.2018.09.008

Swedberg, R. (2018). How to use Max Weber's ideal type in sociological analysis. Journal of Classical Sociology, 18(3), 181–196. https://doi.org/10.1177/1468795X17743643

Taddeo, M. & Floridi, L. (2018). How AI can be a force for good. Science 361:6404, 751–752. doi: 10.1126/science.aat5991

Taeihagh, A. (2021). Governance of artificial intelligence. Policy and Society 40:2, 137–157. doi: 10.1080/14494035.2021.1928377

Taeihagh, A., Ramesh, M. & Howlett, M. (2021). Assessing the regulatory challenges of emerging disruptive technologies. Regulation & Governance 15:4, 1009–1010. doi: 10.1111/rego.12392

The AI Act. (2022). https://artificialintelligenceact.eu/. [accessed 30.11.2022]

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., ... & Clopath, C. (2020). AI for social good: Unlocking the opportunity for positive impact. Nature Communications, 11(1), 1–6.

Torfing, J., Andersen, L., Greve, C. & Klausen, K. (2020). Public governance paradigms: Competing and co-existing. Cheltenham: Edward Elgar Publishing. doi: 10.4337/9781788971225

Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. Human Resource Development Review 15(4), 404–428.

Trajtenberg, M. (2018). Artificial intelligence as the next GPT: A political-economy perspective. In The economics of artificial intelligence: An agenda (pp. 175-186). University of Chicago Press.

Truby, J. (2020). Governing artificial intelligence to benefit the UN sustainable development goals. Sustainable Development 28(4), 946–959. doi: 10.1002/sd.2048

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M. & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. AI & Society. doi: 10.1007/s00146-021-01154-8

Ulnicane, I., Eke, D. O., Knight, W., Ogoh, G. & Stahl, B. C. (2021). Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. Interdisciplinary Science Reviews 46(1-2), 71–93. doi: 10.1080/03080188.2020.1840220

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C. & Wanjiku, W. G. (2020). Framing governance for a contested emerging technology: Insights from AI policy. Policy and Society 40(2), 158–177. doi: 10.1080/14494035.2020.1855800

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Retrieved March 1, 2023, from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Van Roy, V., Rossetti, F., Perset, K. & Galindo-Romero, L. (2021) AI Watch - National strategies on Artificial Intelligence: A European perspective, 2021 edition. EUR 30745 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-39081-7, doi:10.2760/069178, JRC122684

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M. & Nerini, F (2020). The role of artificial intelligence in achieving the sustainable development goals. Natural Communications 11. doi: 10.1038/s41467-019-14108-y

Wallach, W. & Marchant, G. (2018). An agile ethical/legal model for the international and national governance of AI and robotics. Association for the Advancement of Artificial Intelligence.

Wamba, S., Bawack, E. R., Guthrie, C., Queiroz, M. M. & Carillo, K. (2021). Are we preparing for a good AI society? A bibliometric review and research agenda. Technological Forecasting and Social Change 164, 120482. doi: 10.1016/j.techfore.2020.120482

Winfield, A. F. & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2133): 20180085. doi: /10.1098/rsta.2018.0085

Wirtz, B. W., Weyerer, J. C. & Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated AI governance framework for public administration. International Journal of Public Administration 43(9), 818–829. doi: 10.1080/01900692.2020.1749851

Yeung, K., Howes, A. & Pogrebna, G. (2019). “AI governance by human rights-centred design, deliberation and oversight: An end to ethics sashing,” in The Oxford Handbook of AI Ethics, ed. M. Dubber and F. Pasquale. New York: Oxford University Press.

Zicari, R., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., ... & Wurth, R. (2022). How to Assess Trustworthy AI in Practice. arXiv preprint arXiv:2206.09887.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. New York: Public Affairs.

Zuiderwijk, A., Chen, and Y. C. & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly 38(2), 101577. doi: 10.1016/j.giq.2021.101577

Liite A: Menetelmät

Tutkimus toteutettiin systemaattisen integroivan kirjallisuuskatsauksen periaatteita noudattaen. Integroivassa kirjallisuuskatsauksessa tarkastellaan kriittisesti tutkimustietoa ja syntetisoidaan ja integroidaan se uusien mallien tai kehysten luomiseksi (Cronin & George 2020; Torraco 2016). Systemaattisella kirjallisuuskatsauksella voi olla useita eri tavoitteita (Okoli 2015). Tämän katsauksen tavoitteena on edistää tutkimusta, joka tukee tekoälyn eettisen ja vastuullisen käytön julkisen ohjauksen kehittämistä. Tutkimusmenetelmät on kuvattu seikkaperäisemmin Sigfrids ja kumppaneiden (2022) tutkimusartikkelissa.

Tutkimusta varten paikansimme vuosina 2010–2021 (huhtikuu) julkaistuja tekoälyn hallintaa käsitteleviä julkaisuja Taulukossa 1 mainituista tutkimustietokannoista ja siinä mainituilla hakulausekkeilla (Taulukko 1).

Taulukko 6. Tutkimustietokannat ja hakulausekkeet

Tietokannat	Hakulauseke
Web of Science, Scopus, ScienceDirect, Wiley, IEEE **Ebsco, Sage ja Emerald**	("Artificial intelligence" OR “AI” OR "machine learning" OR "deep learning" OR "cognitive computing" OR "artificial neural networks") AND (governance OR "public sector" OR "public administration" OR "government policy")

Siirsimme hakutulokset (n=3392) Mendeley-viitteidenhallintaohjelmaan, joka poisti kaksoiskappaleet. Tulokseksi jäi 2240 artikkelia. Artikkelien määrän rajaamiseksi hyödynsimme kolmivaiheista valintaprosessia. Ensimmäisessä vaiheessa arvioimme tutkimusartikkelien relevanssia aiheemme kannalta. Koska otanta sisälsi suuren määrän selkeästi aiheettomia artikkeleita, yksi tutkija rajasi artikkelit otsikon ja tiivistelmän perusteella niihin, jotka eivät selvästikään soveltuneet mukaan ja niihin, jotka mahdollisesti soveltuivat. Ensimmäisen rajauksen jälkeen tutkimusryhmä keskusteli rajauksen ulkopuolelle jääneistä sekä valituista artikkeleista, minkä jälkeen toteutettiin vielä toinen rajauskierros. Kahden kriittisen lukukierroksen jälkeen tutkimuksen kannalta relevantteja artikkeleita oli 146. Artikkelit valittiin seuraavien kriteerien perusteella:

1. Artikkeli käsittelee tekoälyä.
2. Artikkeli käsittelee ohjausta ja koordinaatiota (eng. governance), etiikkaa tai vastuullisuutta julkishallinnon kontekstissa tai kontekstia määrittelemättä.
3. Artikkelissa tulee olla selkeä yhteiskunnallinen näkökulma eikä artikkeli saa olla pelkästään tai liian tekninen.

Toisessa vaiheessa kaksi tutkijaa arvioivat itsenäisesti abstraktit alla lueteltuja poissulkemissääntöjä noudattaen. Mahdollisista mielipide-eroista keskusteltiin erillisissä kokouksissa. Tutkijat päätyivät 63 artikkeliin. Poissulkemissäännöt olivat seuraavia:

1. Artikkelissa käsitellään pääasiallisesti tekoälyä. Sellaisia tutkimuksia, joissa tekoäly on vain pienessä roolissa, poistettiin.

2. Artikkelin tulisi käsitellä pääasiassa ohjausta ja koordinaatiota, etiikkaa tai vastuullisuutta julkishallinnon yhteydessä tai määrittelemättä kontekstia. Tutkimukset, joissa ohjauksella, etiikalla tai vastuullisuudella oli vain vähäinen merkitys, poistettiin, samoin kuin tutkimukset, jotka koskivat organisaatiokohtaista ohjausta tai tietohallintoa. Algoritmien hyödyntämiseen ohjauksessa, sekä sähköiseen hallintoon, datahallintoon, yritysjohtamiseen tai julkisen sektorin organisaatiotason hallintoon liittyvät tutkimukset poistettiin.

3. Artikkeli ehdottaa tapoja kehittää tekoälyn ohjausta ja koordinaatiota. Sellaiset artikkelit, joissa ei selkeästi esitellä tekoälyn ohjauskeinojen kehittämistä, poistettiin.

Kolmannessa vaiheessa kaksi tutkijaa luki itsenäisesti valittujen artikkelien sisällön arvioidakseen poissulkukriteerien täyttymistä ja kirjoitusten laatua. Näin päädyttiin 21 artikkelin otantaan. Laadun arvioinnissa lähdettiin siitä, että käsitteitä ”governance” ja tekoäly käytetään kirjallisuudessa väljästi ja ne saattavat esiintyä otsikossa tai tiivistelmässä, vaikka niiden osuus sisällöstä on vähäinen tai merkityksetön (ks. Zuiderwijk ym. 2021; Asaduzzaman & Virtanen 2016). Tutkimusten laatua arvioitiin sen perusteella, miten merkityksellisiä käsitteet "governance" ja "tekoäly" olivat tutkimuksen kokonaisuuden kannalta, sekä julkaisumuodon perusteella. Aineiston kerääminen lopetettiin 9.4.2021, minkä jälkeen ilmestyneitä tutkimusartikkeleja ei ole sisällytetty tähän tutkimukseen.

Seuraavaksi valitut artikkelit analysoitiin laadullisen temaattisen analyysin avulla Nowellin ym. (2017) ehdottamien vaiheiden mukaisesti. Tavoitteena oli tunnistaa ja analysoida kaikki tekoälyn ohjauksen kannalta tärkeiksi määritellyt teemat ja ulottuvuudet. Analyysi eteni ja käsitys teemoista muodostui iteratiivisesti usean artikkeleiden lukukierrosten myötä. Analyysin ensimmäisissä vaiheissa tutkijat tutustuivat aineistoon ja loivat alustavia kategorioita jatkoanalyysia varten. Kunkin artikkelin osalta kaksi tutkijaa luki koko artikkelin ja teki samalla muistiinpanoja ennalta määriteltyjen luokkien mukaisesti. Näihin luokkiin kuuluivat perustiedot, kuten viittaus ja tiivistelmä, artikkelissa käsitelty teknologia, tavoitteet, haasteet ja kehitysehdotukset sekä muut huomiot. Muistiinpanot koottiin alustaviksi teemoiksi, joita kehiteltiin edelleen lukemalla artikkelit iteratiivisesti uudelleen seuraavissa vaiheissa. Uudelleenlukemisessa kiinnitettiin erityistä huomiota seuraaviin seikkoihin: (1) Ongelman määrittelyyn eli millä tavalla tekoälyn ohjaus esitetään kehittämistä vaativana aiheena. Nämä olivat tyypillisesti kuvauksia tekoälyyn liittyvistä kysymyksistä, jotka kirjoittajan mukaan edellyttävät ohjauskeinojen kehittämistä riippumatta olemassa olevista käytännöistä ja lainsäädännöstä. (2)

Kuvauksiin sellaisista periaatteista ja tavoitteista, joihin nojaten tekoälyn ohjausta tulisi kehittää. (3) Kuvauksiin keinoista ja menettelyistä, joita tarvitaan tekoälyn ohjauksen kehittämiseksi. Analyysin viimeisissä vaiheissa teemat nimettiin ja yhdistettiin ideaalityypeiksi, jotka koostuvat kirjallisuudessa esitettyjen ohjausehdotusten keskeisistä johtopäätöksistä.

Julkaisun sarja ja numero

VTT Technology 421

Nimeke	**Julkishallinto tukemassa eettisen tekoäly-yhteiskunnan rakentamista** Katsaus tekoälyn ohjauskeinoihin ja politiikkatoimenpiteisiin
Tekijä(t)	Anton Sigfrids, Mika Nieminen, Jaana Leikas, Antero Karvonen & Pietari Pikkuaho
Tiivistelmä	Tekoälyn edistysaskeleet herättävät kysymyksiä tekoälysovellusten yhteiskunnallisista vaikutuksista. Monet valtiot, yritykset ja kansainväliset organisaatiot ovat kehittäneet ohjauskeinoja, joilla on pyritty lieventämään tekoälyyn liittyviä riskejä ja maksimoimaan sen hyödyntämisen tuomat edut. Ohjaus- ja koordinaatiokeinoja kehitettäessä keskeisiksi kysymyksiksi nousevat miten haasteita ja riskejä olisi hallittava, millaisten arvojen perusteella toimitaan, millaisia tavoitteita asetetaan ja millaisten institutionaalisten mekanismien ja periaatteiden avulla tavoitteet voidaan saavuttaa. Tekoälyn ohjaus- ja koordinaatiokeinot kehittyvät nopeasti kansallisilla ja kansainvälisillä foorumeilla ja myös tutkimuskirjallisuutta julkaistaan kiihtyvään tahtiin. Tekoälyn hallinta etsii muotoaan. Tässä raportissa koostamme ja tarkastelemme tutkimuskirjallisuudessa esiintyviä ehdotuksia julkisen hallinnon ohjaus- ja koordinaatiomekanismien parantamiseksi. Kiinnitämme erityistä huomiota sellaisiin periaatteisiin ja keinoihin, joiden avulla julkishallinto voi ohjata tekoälyn kehittäjiä ja käyttäjiä omaksumaan eettisiä ja vastuullisia käytäntöjä. Tarkastelumme osoittaa, että julkishallinnon tulisi omaksua osallistavan päätöksenteon muotoja pyrkimyksissään kokonaisvaltaisempaan ja koordinoituun tekoälyn ohjaukseen, valvontaan, sekä käyttötilanteittain räätälöityihin eettisiin toimintamalleihin. Ehdotamme ratkaisuksi ideaalityyppistä OSKI-mallia (Osallistava, Soveltuva, Kokonaisvaltainen ja Institutionalisoitu), jossa ehdotettujen kehittämisratkaisujen keskeiset ulottuvuudet yhdistyvät tekoälyn käytön kokonaisvaltaiseksi ohjausmalliksi. Tämän tekstin aiempi versio on julkaistu englanninkielisenä artikkelina Sigfrids ym. (2022). Raportti on toteutettu strategisen tutkimuksen neuvoston rahoittamassa ETAIROS-tutkimushankkeessa, jonka keskeinen tavoite on vahvistaa julkisen ja yksityisen sektorin yhteistä ymmärrystä tekoälykehityksen eettisistä käytännöistä ja pelisäännöistä. ETAIROS kehittää eettisesti kestäviä tekoälyn suunnittelumenetelmiä ja tuottaa näkemystä tekoälyä koskevan koordinaation ja ohjauksen keinoista yhdessä sidosryhmien kanssa.
ISBN, ISSN, URN	ISBN 978-951-38-8784-1 ISSN-L 2242-1211 ISSN 2242-122X (Verkkojulkaisu) DOI: 10.32040/2242-122X.2023.T421
Julkaisuaika	Heinäkuu 2023
Kieli	Suomi
Sivumäärä	54 s. + liitt. 3 s.
Projektin nimi	ETAIROS
Rahoittajat	
Avainsanat	Tekoäly, tekoälyn etiikka, tekoälyn ohjaus, vastuullinen tekoäly, julkinen ohjaus
Julkaisija	Teknologian tutkimuskeskus VTT Oy PL 1000, 02044 VTT, puh. 020 722 111, https://www.vtt.fi/

Julkishallinto tukemassa eettisen tekoäly-yhteiskunnan rakentamista

Katsaus tekoälyn ohjauskeinoihin ja politiikkatoimenpiteisiin

Tekoälystä on niin Suomessa kuin kansainvälisesti tulossa yleiskäyttöteknologia, jonka nähdään tuottavan hyötyjä monella eri alalla. Tekoälysovelluksiin liittyy kuitenkin tahattomia ja ristiriitaisia vaikutuksia, jotka voivat olla haitallisia yksilöille tai yhteiskunnalle. Nykyiset julkiset ohjauskeinot eivät ehkäise tekoälyn käytön nostattamia sosiaalisia riskejä ja uhkia riittävästi, minkä vuoksi niitä tulee kehittää.

Organisaatioilla ja julkishallinnolla tulisi olla kyvykkyyttä ennakoida ja tunnistaa tekoälyn riskejä kokonaisvaltaisesti ja vastata ketterästi syntyviin ongelmiin. Onnistuessaan julkinen ohjaus luotsaa tekoälyn kehittäjiä ja käyttäjiä siten, että tekoäly tuottaa yhteistä hyvää niin yksilöiden, yhteisöjen kuin koko yhteiskunnan kannalta. Tämä tapahtuu koordinoimalla politiikkatoimia nykyistä selvemmin sekä tukemalla vastuullista toimintaa ja kyvykkyyttä hahmottaa tekoälyn eettisiä vaikutuksia systeemisesti.

Esittelemme tässä raportissa systemaattisen kirjallisuuskatsauksen perusteella (Sigfrids ym. 2022) laatimamme tekoälyn julkisen ohjauksen tukemiseksi tarkoitetun OSKI-mallin (Osallistava, Soveltuva, Kokonaisvaltainen ja Institutionalisoitu). OSKI kuvaa ideaalitilaa ja tarjoaa julkisohjauksen suunnittelulle ja kehittämiselle suuntaviivat, joita kohti pyrkiä.

ISBN 978-951-38-8784-1
ISSN-L 2242-1211
ISSN 2242-122X (Verkkojulkaisu)
DOI: 10.32040/2242-122X.2023.T421

VTT

beyond the obvious