TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ALGORITMISET HAITAT

Tekoälyn riskit ja sääntelyn haasteet kiihtyvässä yhteiskunnassa

Nea Lepinkäinen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 675 | HUMANIORA | TURKU 2024

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ALGORITMISET HAITAT

Tekoälyn riskit ja sääntelyn haasteet kiihtyvässä yhteiskunnassa

Nea Lepinkäinen

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. B OSA – TOM. 675 | HUMANIORA | TURKU 2024

Turun yliopisto

Oikeustieteellinen tiedekunta
Oikeustieteen tohtoriohjelma

Työn ohjaajat

Anne Alvesalo-Kuusi
Oikeustieteen professori
Turun yliopisto

Hanna Malik
Dosentti (vertaileva oikeustiede ja oikeussosiologia)
Turun yliopisto

Mika Viljanen
Oikeustieteen professori
Turun yliopisto

Tarkastajat

Riikka Koulu
Apulaisprofessori (tekoälyn yhteiskunnallis-oikeudelliset vaikutukset)
Helsingin yliopisto

Terhi Esko
FT, tutkijatohtori
Helsingin yliopisto

Vastaväittäjä

Riikka Koulu
Apulaisprofessori (tekoälyn yhteiskunnallis-oikeudelliset vaikutukset)
Helsingin yliopisto

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä.

ISBN 978-951-29-9824-1 (painettu)
ISBN 978-951-29-9825-8 (verkko)
ISSN 0082-6987 (painettu)
ISSN 2343-3191 (verkko)
Painosalama, Turku, Suomi 2024

TURUN YLIOPISTO
Oikeustieteellinen tiedekunta
NEA LEPINKÄINEN: Algoritmiset haitat: Tekoälyn riskit ja sääntelyn haasteet kiihtyvässä yhteiskunnassa
Väitöskirja, 236 sivua (sisältäen 3 alkuperäisjulkaisua)
Oikeustieteen tohtoriohjelma
Elokuu 2024.

TIIVISTELMÄ

Tekoäly on levittänyt vaikutustaan yhä laajemmin yhteiskunnan eri osa-alueille, ja algoritmisen transformaation vauhti näyttäisi vain kiihtyvän. Algoritmiset teknologiat ovat kietoutuneet erottamattomaksi osaksi yhteiskuntaa, ja niiden mukana yhteiskunnassa ovat yleistyneet myös algoritmiset haitat.

Tässä artikkeliväitöskirjassa keskiössä ovat algoritmisten teknologioiden aiheuttamat moninaiset haitat, haittojen muodostumiseen vaikuttavat tekijät sekä tekoälysääntelyn tarjoamat mahdollisuudet haittojen hallitsemiseen. Vaikka osa tekoälyn vaikutuksista on positiivisia, väitöskirjassa keskitytään nimenomaan potentiaalisiin ja jo tunnettuihin algoritmisiin haittoihin. Tutkimus kiinnittyy erityisesti yhteiskunnallisten haittojen (social harm) tutkimusperinteeseen. Modernin yhteiskunnan kehityskulkujen, etenkin tekoälyn alati lisääntyvän käytön ymmärtämiseksi ja analysoimiseksi hyödynnetään Rosan yhteiskunnallisen kiihtyvyyden teorian (theory of social acceleration) ja Giddensin rakenteistumisteorian (structuration theory) teesejä.

Väitöskirjassa arvioidaan, minkälaisia algoritmisia haittoja tekoälyteknologiat aiheuttavat tai voivat aiheuttaa nyky-yhteiskunnassa. Työ laajentaa yhteiskunnallisten haittojen teorian näkökulmaa ottamalla huomioon yksilöihin ja ryhmiin kohdistuvien haittojen lisäksi haitat, jotka kohdistuvat yhteiskuntaan. Tekoälyteknologioiden teknisistä ominaisuuksista juontuvien, haitoille altistavien riskien paremmaksi ymmärtämiseksi väitöskirjassa tukeudutaan Viljasen (2022) hahmottelemaan tekoälyn teknisten ominaisuuksien luokitteluun.

Haittojen ja tekoälyn ominaisuuksien vaikutusten analyysi paljastaa, että vasta osa algoritmisiin teknologioihin liittyvästä haittapotentiaalista on tunnistettu. Etenkin yhteiskuntaan kohdistuvia algoritmisia haittoja on tutkittu vasta vähän. Kansallisen ja EU-tason tekoälysääntelyratkaisujen jäsentely osoittaa, että myös lainsäädännössä onnistutaan huomioimaan ainoastaan osa haittapotentiaalista, joka algoritmisiin järjestelmiin liittyy. Tuoreet lainsäädäntöratkaisut pyrkivät löytämään tasapainon siten, että tekoälyn mukanaan tuomat riskit saataisiin hallittua ilman liian suuria rajoituksia innovaatioille tai kilpailulle. Algoritmisten teknologioiden tuotantoa ohjaava, mahdollisimman vähän markkinoita rajoittava riskiperustainen EU-sääntely tunnistaa ainoastaan välittömät haitat. Sen sijaan se ei vaikuta tunnistavan kiihtyvään algoritmiseen transformaatioon liittyviä seurannaishaittoja, jotka uhkaavat perustavanlaatuisella tavalla sekä yhteiskunnan toiminnan että inhimillisen kukoistuksen mahdollisuuksia.

Tässä väitöskirjassa argumentoidaan, että algoritmista transformaatiota seuraavat haitat olisi mahdollista tuoda valittuja sääntelyratkaisuja tehokkaammin säänte-

lyn piiriin, mikäli poliittista tahtoa tällaiseen olisi. Yksi keino tämän toteuttamiseksi olisi laajentaa EU:n tekoälysääntelyn vaikutusarviointimenettely koskemaan kaikkia tekoälysovelluksia ja sitoa se vahvemmin nimenomaan potentiaalisesti seuraavien haittojen arviointiin. Tästä mahdollisesti seuraava tekoälyteknologioiden kehityksen hidastuminen olisi mahdollista nähdä neutraalina seikkana tai jopa tavoitteena, ei ongelmana: se lisäisi mahdollisuuksia hillitä algoritmisen transformaation vauhtia ja suunnata yhteiskunnan muutosta kohti tulevaisuutta, jossa inhimillinen kukoistus turvattaisiin yhä laajemmin. Tämän sijaan sääntelyssä kuitenkin pyritään välttämään laajoja rajoituksia markkinoiden ja innovaatiokehityksen turvaamiseksi. Ratkaisun riskinä on, että ennalta estämisen sijaan entistä useammin eurooppalaiset yhteiskunnat joutuvat jälkikäteen ratkomaan algoritmisten teknologioiden kiihtyvästä kehityksestä ja käytöstä sekä algoritmisesta transformaatiosta juontuvia moninaisia algoritmisia haittoja.

ASIASANAT: algoritmiset teknologiat, tekoäly, yhteiskunnalliset haitat, tekoälysääntely, yhteiskunnallinen kiihtyminen, algoritminen transformaatio

UNIVERSITY OF TURKU
Faculty of Law
NEA LEPINKÄINEN: Algorithmic Harms: AI Risks and Regulatory Challenges in an Accelerating Society
Doctoral Dissertation, 236 pp. (including 3 original publications)
Doctoral Programme in Law
August 2024

ABSTRACT

Artificial intelligence has increasingly spread its influence across various sectors of society, and the pace of algorithmic transformation seems to be accelerating. Algorithmic technologies have become an integral part of society, and as they have spread, so too have algorithmic harms.

This article-based dissertation focuses on the diverse harms caused by algorithmic technologies, the factors influencing the formation of these harms, and the potential offered by AI regulation for managing them. While some effects of AI are positive, the dissertation specifically addresses the potential and already known algorithmic harms. The research is particularly anchored in the tradition of social harm studies. To understand and analyse the developments in modern society, especially the increasing use of AI, the theses of Rosa's theory of social acceleration and Giddens' structuration theory are employed.

This dissertation assesses different types of algorithmic harms that AI technologies cause or could potentially cause in contemporary society. The work broadens the perspective of social harm theory by considering harms that affect society in addition to those impacting individuals and groups. To better understand the risks inherent in the technical characteristics of AI technologies that predispose them to cause harm, the dissertation relies on Viljanen's (2022) classification of the technical characteristics of AI.

The analysis of harms and the impact of AI characteristics reveals that only a portion of the potential harms associated with algorithmic technologies has been identified. In particular, algorithmic harms affecting society have been relatively under-researched. The review of national and EU-level AI regulation solutions indicates that legislation only manages to consider part of the potential harms related to algorithmic systems. Recent legislative solutions aim to strike balance by managing the risks associated with AI without imposing excessive restrictions on innovation or competition. The EU's risk-based regulation that governs the production of algorithmic technologies, while minimally restricting markets, only identifies immediate harms. However, it does not seem to recognise the consequential harms associated with the accelerating algorithmic transformation, which fundamentally threaten both the functioning of society and the potential for human flourishing.

This dissertation argues that the harms following algorithmic transformation could be more effectively brought under regulatory control, provided there is political will for such action. One way to achieve this would be to expand the impact assessment procedure of EU AI Act to cover all AI applications, with a stronger

focus specifically on assessing potential consequential harms. The potential slowdown in the development of AI technologies that might result from this could be seen as a neutral aspect or even a goal, rather than a problem: it would increase the opportunity to moderate the pace of algorithmic transformation and steer societal change towards a future where human flourishing is more broadly secured. However, instead of this, the current regulatory approach seeks to avoid broad restrictions to safeguard markets and innovation development. The risk of this approach is that, instead of preventing harms beforehand, European societies may increasingly find themselves having to address a growing number of algorithmic harms, stemming from the accelerated development and use of algorithmic technologies, and the resulting algorithmic transformation, after they occur.

KEYWORDS: algorithmic technologies, artificial intelligence, social harm, AI regulation, social acceleration, algorithmic transformation

Kiitokset

Aloitin väitöskirjani työstämisen vuonna 2018. Tämän viisivuotisen projektin aikana en ole ollut joka hetki aivan varma, että projekti tulisi valmiiksi, vaan joukossa on ollut päiviä, kun sana ”valmis” on tuntunut kuuluvan puhtaasti satumaailmoihin ja mahdottomiin haaveisiin. Siitä huolimatta tässä sitä ollaan: väitöskirjan kiitoksien parissa, viimeisissä hetkissä ennen tämän tutkimuksen laskemista lopullisesti käsistä. Yksin en olisi päässyt tänne asti, ja haluankin käyttää tämän tilaisuuden kiittääkseni kaikkia niitä, jotka ovat tukeneet minua ja jaksaneet uskoa minuun ja tutkimukseeni silloinkin, kun oma uskoni on ollut vähissä.

Tekoälyyn liittyvät oikeudelliset kysymykset eivät olleet ensimmäinen ideani väitöskirjan tutkimusaiheeksi, eihän aihepiiri vuonna 2017 tohtorikoulutukseen hakeutumista miettiessäni ollut saanut vielä lainkaan nykyisenkaltaista huomiota. Sen sijaan ajatuksenani oli pyrkiä tutkimaan huumeisiin liittyviä kysymyksiä. Tekoäly kuitenkin kiehtoi, ja siihen liittyvät teemat nousivat esille tutkimussuunnitelmaa kehitellessäni yhä uudestaan ja yhä mielenkiintoisemmissa yhteyksissä. Lisäksi minua suorastaan yllytettiin hakeutumaan tohtorikoulutukseen juuri tekoälyyn liittyvien oikeudellisten teemojen tutkimiseksi. Niinpä hylkäsin huumeet ja siirryin tekoälyn pariin, kirjoitin tutkimussuunnitelman ja hain tohtorikoulutettavaksi. Te minua tähän ratkaisuun yllyttäneet, tiedätte keitä olette, ja olen teille syvästi kiitollinen.

Tästä lähtökohdasta väitöskirjatyöni tekemisen on mahdollistanut Turun yliopiston oikeustieteellinen tiedekunta. Olen viettänyt tiedekunnassa aikaani enemmän tai vähemmän tiiviisti vuodesta 2011 lähtien, siitä viimeiset viisi vuotta pääosin väitöskirjani parissa, ja saanut oppia elämästä, tieteestä ja akateemisen maailman jännittävistä kiemuroista. Ilman tiedekunnan tarjoamaa valtaisaa tiedollista, henkistä, fyysistä – ja tietenkin myös rahallista – tukea tämä työ ei olisi koskaan päässyt alkamaan tai valmis työ siintäisi vasta kaukana horisontissa. Rahallista tukea ovat tarjonneet myös Suomen Akatemian projektit ETAIROS ja AALAW, mistä olen valtavan kiitollinen.

Tutkimustyö ei ole helppoa, kuten varmasti jokainen tutkija tietää. Akateemisen maailman ulkopuolelta tutkijuuteen kasvaminen ei ole ollut minulle lainkaan kivutonta, ja paikoin kasvukipujen voimakkuus on päässyt jopa yllättämään. Ajattelun kehittäminen, argumentaatiotaitojen vahvistaminen ja oman kriittisen äänen harjaan-

nuttaminen jatkuvat tietenkin tästä eteenpäinkin, mutta ensimmäiset askeleet niiden saralla on nyt otettu. En olisi päässyt näin pitkälle ilman ohjaajieni opastusta. Mika Viljanen ja Anne Alvesalo-Kuusi, teidän asiantuntemuksenne sekä tarjoamanne tuki ovat olleet korvaamattomia tällä oppimatkalla ja mahdollistaneet työni loppuun saattamisen. Hanna Malik, kanssasi tehty yhteistyö on opettanut minulle suurimman osan siitä, mitä akateemisesta tutkimuksesta ymmärrän. Kiitos, että lähdit ennakkoluulottomasti kirjoittamaan kanssani tähän työhön sisältyviä osatutkimuksia ja tarjosit minulle paitsi ohjausta myös ystävyyttä. Se on ollut minulle äärimmäisen arvokasta.

Vaikka olen ollut erittäin onnellisessa asemassa saadessani kirjoittaa tämän väitöskirjan osatutkimukset yhteistyössä todella osaavien ja kokeneiden tutkijoiden kanssa, on tutkimustyö silti monin paikoin ollut hyvin itsenäistä ja osin myös yksinäistä. Omaa yksinäisyyden kokemustani varmasti voimisti myös väitöskirjatyöni alkupuolella maailmaa ravisuttanut koronapandemia, jonka myötä etätöistä tuli uusi normaali, konferenssit ja seminaarit peruttiin tai siirrettiin nettiin, ja ihmisjoukot muuttuivat virkistävistä uhkaaviksi. En tiedä, kuinka olisin selvinnyt ilman läheisiäni, joiden pyyteetön tuki, kannustus ja kärsivällisyys ovat nostaneet minut monista alhoista, ja jotka ovat jaksaneet kiskoa minut tekstien ääreltä mukaan myös fyysisen maailman menoihin virkistämään ajatuksiani. Kumppanini, perheeni, ystäväni, kollegani: ilman teitä en olisi nyt tässä.

Lopuksi haluan kiittää väitöskirjani esitarkastajia Riikka Koulua ja Terhi Eskoa. Teidän perusteellinen palautteenne ja rakentavat ehdotuksenne työni parantamiseksi auttoivat minua paitsi viimeistelemään työni myös tarkastelemaan ja syventämään omia näkemyksiäni entisestään. Kiitos, ilman teidän panostanne tämä työ ei olisi yltänyt tälle tasolle.

05.08.2024
Nea Lepinkäinen

Sisällys

Osajulkaisuluettelo

Väitöskirjan yhteenveto-osa perustuu seuraaviin alkuperäisjulkaisuihin.

1 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, 11(1), 182–195.
2 Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, 3(1), 33707.
3 Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, 5(3), 270–291.

Artikkelien käyttöön väitöskirjan osajulkaisuina on saatu kustantajien lupa.

1 Tutkimuksen tausta ja teoreettiset lähtökohdat

Tekoäly ja sen käyttö ovat viime vuosina nousseet eri valtioiden poliittisilla agendoilla tärkeiksi teemoiksi. Tekoäly tarjoaa valtavasti mahdollisuuksia uusien toimintojen luomiseen ja olemassa olevien tehostamiseen. Näitä mahdollisuuksia sekä yksityinen että julkinen sektori pyrkivät hyödyntämään niin uusien palveluiden tarjoamisessa kuin olemassa olevien toimintojen tehostamisessa. Algoritmisaation kiihtymistä ja tekoälyn merkityksen kasvamista läpi yhteiskunnan voidaan käsitteellistää jatkumona digitalisaatiolle, joka alkoi jo 1980-luvulla tietokoneiden yleistymisen vanavedessä, ja jonka myötä tietotekniset ratkaisut ja digitaaliset teknologiat ovat jatkuvasti lisänneet merkitystään moderneissa yhteiskunnissa[1]. Viime vuosina tämä kehitys on johtanut siihen, että algoritmiset teknologiat ovat muuntuneet työkaluista ihmisten jokapäiväistä elämää ja toiminnan tapoja muovaaviksi voimiksi. Ne vaikuttavat ihmisten elinympäristöön, sosiaaliseen kanssakäymiseen, tiedon rakentumiseen ja maailmankuvien muodostumiseen. Muutoksen vaikutuksista onkin viime vuosina kirjoitettu paljon. Siinä missä toiset argumentoivat optimistisessa hengessä tekoälyn avaamien mahdollisuuksien puolesta[2], toiset näkevät tekoälyn väistämättä lisäävän eriarvoisuutta[3] ja

1 Dornberger, R., Inglese, T., Korkut, S., & Zhong, V. J. (2018). Digitalization: Yesterday, today and tomorrow. *Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change*, 1–11.

2 Ossewaarde, M., & Gulenc, E. (2020). National varieties of artificial intelligence discourses: Myth, utopianism, and solutionism in West European policy expectations. *Computer*, *53*(11), 53–61;

Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

3 McQuillan, Dan. *Resisting AI: an anti-fascist approach to artificial intelligence*. Policy Press, 2022.

Mann, M. (2020). Technological politics of automated welfare surveillance: Social (and data) justice through critical qualitative inquiry. *Global Perspectives*, *1*(1).

uhkaavan demokratiaa[4], oikeusvaltioperiaatetta[5] ja sosiaalista oikeudenmukaisuutta[6]. Niin kutsuttu supertekoäly on noussut uhkakuvaksi populaarikulttuurin perinteisestä kuvastosta myös tieteen kentälle[7]. Algoritmisten teknologioiden valtavaan muutosvoimaan pohjautuu ilmiö, jota kutsun *algoritmiseksi transformaatioksi*. Siinä missä digitalisaation ja *digitaalisen transformaation* on argumentoitu muuttaneen perustavanlaatuisella tavalla niin liike-elämää, yhteiskunnan toimintoja kuin ihmisten jokapäiväistä arkea[8], algoritminen transformaatio edelleen kiihdyttää, voimistaa ja syventää tätä digitalisaation pohjustamaa muutosta, jonka ennenkuulumattoman laajat vaikutukset – niin positiiviset kuin negatiivisetkin – ilmenevät vaihtelevalla intensiteetillä ja epätasaisesti eri yhteiskuntien, yhteiskuntaluokkien, ryhmien ja yksilöiden jokapäiväisessä todellisuudessa.

Tekoälyn ja algoritmisaation merkityksestä kertovat myös laajat pyrkimykset teknologioiden ja/tai niiden käyttämisen sääntelyyn. Suomessa on vuoden 2023 toukokuun alusta lähtien mahdollistettu automaattinen päätöksenteko julkishallinnossa tietyin edellytyksin. EU:ssa astui voimaan elokuussa 2024 tekoälysäädös ((EU) 2024/168), joka siirtymäaikojen jälkeen velvoittaa EU:n jäsenvaltioita sellaisenaan. Tekoälysäädöksen tarkoituksena on tukea ihmiskeskeisen tekoälyn kehitystä ja käyttöönottoa, kannustaa innovaatioihin ja varmistaa, että tekoälytuotteet eivät EU:n alueella vaaranna ihmisten terveyttä, turvallisuutta tai ihmisoikeuksien toteutumista (artikla 1). Osaltaan tekoälysäädöksellä pyritään siis hallitsemaan riskejä, joita algoritmisten teknologioiden on havaittu aiheuttavan. Tätä tavoitetta kuvastaa säädöksessä omaksuttu ratkaisu jakaa tekoälyjärjestelmät niiden käyttötarkoituksen, käyttökontekstin ja teknisten ratkaisujen perusteella eri riskiluokkiin, joihin liittyy erilaisia velvoitteita järjestelmien tuottajille ja käyttöönottajille.

Tekoälystä käytävään keskusteluun liittyy huomattava määrä termejä, joita aiheeseen perehtyneenkin voi toisinaan olla vaikea ymmärtää. Käytän tässä työssä ter-

4 Yeung, K. (2011). Can we employ design-based regulation while avoiding brave new world? *Law, Innovation and Technology*, *3*(1), 1–29.

5 Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review*, *82*(3), 425–455.

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128), 20170355.

6 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

7 Esimerkiksi Bostrom, N. (2014). *Superintelligence*. Oxford University Press;

Tegmark, M. (2018). Life 3.0: Being human in the age of artificial intelligence. Vintage.

8 Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *Sage Open*, *11*(3), 21582440211047576.

mejä algoritmiset teknologiat, algoritmiset järjestelmät sekä tekoälyteknologiat ja -järjestelmät pitkälti synonyymeinä, joskin tekoäly-termiin sisältyy näistä voimakkaampi painotus järjestelmätasoiselle monimutkaisuudelle. Molemmilla termeillä kuitenkin viitataan teknologioihin, jotka pystyvät jokseenkin itsenäisesti tuottamaan ratkaisuja monimutkaisiin pulmiin joko opetusdataan perustuvien ennustemallien avulla (koneoppivat järjestelmät[9], *machine learning systems*) tai tietokantojen ja sääntöjen pohjalta päättelemällä ja ennustamalla (tieto- ja logiikkapohjaiset järjestelmät, *knowledge and logic based systems*). Tällaiset teknologiat näen väistämättä eettispoliittisina järjestelminä[10], jotka kiinnittyvät osaksi nykymaailman sosioekonoteknistä[11] systeemiä. Ne siis vaikuttavat sekä yhteiskunnan rakenteellis-toiminnallisiin järjestelmiin että ihmisten, koneiden ja ympäristötekijöiden muodostamiin sosioteknisiin järjestelmiin, ja niiden toimintaa ohjaavat yhtäältä poliittiset, toisaalta eettiset ratkaisut, joita järjestelmien suunnittelun, tuotannon ja käyttökohteiden valinnoissa tehdään.

Käytännössä algoritmisia teknologioita on äärimmäisen monenlaisia. Kuten aiemmin olemme yhdessä kollegoideni kanssa esittäneet:

> As layered, multifaceted tools, these technologies cannot be compressed into a single phenomenon. Any attempt to do so would ultimately fail to consider the countless possibilities, risks and contradictions provided by different technological solutions. This is important from a policy and regulatory perspective, even though quite ironically, extensive evidence reveals that the use of automatically enhanced classification and decision-making systems reduces the complexity of life itself.[12]

Tämä tekee tekoälyteknologioiden tutkimisesta haastavaa. Erilaisilla teknologoilla on vaihtelevasti potentiaalia vaikuttaa paitsi ihmisiin ja näiden ajatteluun ja toimintaan, myös laajemmin ympäristöönsä ja yhteiskunnan toimintoja määrittäviin rakenteisiin. Erilaisten teknologioiden tekniset ominaisuudet, käyttöympäristö ja

9 Koneoppivat järjestelmät viittaavat järjestelmiin, jotka on koulutettu koneoppimismenetelmien avulla. Jokseenkin yleinen, mutta virheellinen mielikuva on, että koneoppivat järjestelmät oppisivat käytössä ollessaan. Näin on äärimmäisen harvoin. Tässä työssä puhutaan *dynaamisista järjestelmistä*, mikäli tarkoituksena on viitata käytön aikana oppiviin järjestelmiin.

10 Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology*, *47*, 213–237.

11 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

12 Malik, H. M., Lepinkäinen, N., Alvesalo-Kuusi, A., & Viljanen, M. (2022). Social harms in an algorithmic context. *Justice, Power and Resistance*, *5*(3), 193–207.

käytön mahdollisuudet vaihtelevat valtavasti. Algoritmisten teknologioiden vaikutuksia ja haittapotentiaalia analysoitaessa onkin syytä arvioida paitsi, miten tekoälyjärjestelmien ominaisuudet vaikuttavat niiden toimintaan ja toiminnasta mahdollisesti seuraaviin haittoihin, myös, missä ja miten järjestelmiä käytetään ja miten sosioekonotekninen ympäristö, jossa järjestelmät toimivat, mahdollisesti muuttuu teknologioiden vaikutuksesta.

Algoritmisten järjestelmien tutkimus, samoin kuin tutkimus tällaisten järjestelmien ja yhteiskunnan välisistä vuorovaikutussuhteista, on viime vuosina lisääntynyt valtavasti. Tekoälyteknologioiden potentiaalisia ja toteutuneita vaikutuksia on jo arvioitu laajasti muun muassa kriittisen algoritmi- ja tekoälytutkimuksen, STS:n (*science and technology studies*) ja digitaalisen kriminologian (*digital criminology*) parissa. Myös tulevaisuudentutkimuksen kentällä on yhä laajemmin kiinnostuttu tekoälystä ja siihen liittyvistä riskeistä ja mahdollisuuksista. Tekoälyteknologioita tutkivia kriittisiä suuntauksia yhdistää pyrkimys arvioida tekoälyteknologioita osana yhteiskunnan olemassa olevaa sosioteknistä todellisuutta[13]. Tämän todellisuuden osana ilmenevät myös vallan epätasaiseen jakautumiseen kiinnittyvät ilmiöt ja niihin liittyvät haitat, jotka ovat keskiössä yhteiskunnallisten haittojen tutkimuksessa.

Tämän työn tarkoituksena on ensi sijassa tuoda tekoälytutkimuksen kriittisten suuntausten antia vahvemmaksi osaksi *yhteiskunnallisten haittojen (social harm)*[14] tutkimusta, jota kutsutaan myös *zemiologiaksi*[15]. Yhteiskunnallisten haittojen tutkimuksen pääasiallisena tavoitteena on tunnistaa yhteiskunnallisesti tuotettuja haittoja, jotka usein jäävät rikosoikeuskeskeisen valtavirtakriminologian tutkimusteemojen ulkopuolelle. Haittojen tutkimus auttaa paitsi ymmärtämään yhteiskunnallista todellisuutta myös suuntaamaan akateemista ja myös *oikeudellista* kiinnostusta ilmiöihin, joiden tutkimusta tulisi lisätä tai sääntelyä kehittää. Siinä mielessä haittojen tutkimus etääntyy valtavirtakriminologiasta: kriminologit ovat kiinnostuneita erityisesti rikosoikeusjärjestelmästä ja tutkivat siihen liittyviä ilmiöitä, kun taas haittojen tutkimuksessa kiinnostus kohdentuu huomattavasti laajemmalle alueelle. Yhteiskunnallisten haittojen tutkimusperinteessä argumentoidaankin laajasti, että rikoslaki ja kriminalisoinnit eivät riitä haittojen hallitsemisen työkaluiksi. Kriittisen tekoälytutkimuksen valossa näin näyttäisi olevan myös tekoälyjärjestelmien kohdalla. Tämän työn loppupuolella arvioinkin, miten tekoälyn sääntelypyrkimyksissä tunnistetaan tekoälyyn

13 Raley, R., & Rhee, J. (2023). Critical AI: A field in formation. American Literature, 95(2), 185–204.

14 Pemberton, S. (2015). Harmful Societies: Understanding Social Harm. Policy Press; Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (2004). *Beyond criminology: Taking harm seriously*. Pluto.

15 Canning, V., & Tombs, S. (2021). From social harm to zemiology: A critical introduction. Routledge.

liittyvät haitat ja minkälaisia mahdollisuuksia niiden avulla on hallita haittojen ilmenemistä.

Tässä johdanto-osassa jatkan johtopäätelmästä, johon ensimmäisessä osatutkimuksessa[16] päädyttiin: algoritmisilla teknologioilla on merkitystä, kun arvioidaan yhteiskunnallisten haittojen syntydynamiikkaa. Kun teknologiat kiihdyttävät modernin yhteiskunnan prosesseja, uusia haittoja syntyy, tunnettujen haittojen ominaispiirteet ja leviämisen tavat muuttuvat, ja haittojen hallinnasta tulee entistä hankalampaa. Teeman kannalta *aika* ja toisaalta siihen kytkeytyvä *muutos* muodostavat merkittävät parametrit. Yhteiskunnallisten haittojen teorian lisäksi tarvitaankin myös teorioita, joiden kautta näitä parametrejä voidaan ymmärtää ja analysoida teknologisen kehityksen kontekstissa. Yhteiskunnallisen kiihtymisen teoria (theory of social acceleration)[17] auttaa hahmottamaan modernin yhteiskunnan ajallista ulottuvuutta, kun taas rakenteistumisteoria (*structuration theory*)[18] auttaa jäsentämään yhteiskunnan rakenteiden ja toimijoiden välisiä vuorovaikutussuhteita ja niiden roolia yhteiskunnan muutoksissa.

Suomenkielinen haittoihin keskittyvä kirjallisuus on vähäistä, eikä myöskään Rosan yhteiskunnallisen kiihtyvyyden teoriasta ole vielä kirjoitettu suomeksi paljoa. Näin ollen myös terminologia on vakiintumatonta. *Theory of social harm* on käännetty tässä tutkimuksessa yhteiskunnallisten haittojen teoriaksi, ja *theory of social acceleration* yhteiskunnallisen kiihtyvyyden teoriaksi. Suomeksi on kuitenkin löydettävissä jonkin verran aineistoa, jossa *social* on yhteiskunnallisen sijaan käännetty *sosiaaliseksi*, ja näin ollen puhutaan *sosiaalisista haitoista*[19] tai *sosiaalisesta kiihtyvyydestä*[20]. Tämä korostaa vahvasti ihmistenvälisyyttä ja jättää rakenteellisten, systemaattisten ja vallan epätasaiseen jakautumiseen liittyvien seikkojen painoarvon ymmärtämisen lukijan oman valistuneisuuden varaan. *Yhteiskunnallinen* toivoakseni nostaa myös aihepiiriin vähemmän perehtyneelle ihmiselle ensimmäisenä mieleen ajatuksen yhteiskunnallisiin prosesseihin tai yhteiskunnalliseen järjestelmään liittyvistä ilmiöistä. Toisaalta riskinä on, että ihmisen kokemuksen ja toiminnan tai ihmisten välisten suhteiden painoarvo ei nouse yhtä selkeästi esille, jolloin yksilöihin tai ryhmiin kohdistuvat yhteiskunnalliset haitat voidaan virheellisesti mieltää toissijaisiksi.

16 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. International Journal for Crime, Justice and Social Democracy, 11(1), 182–195.

17 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

18 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

19 Ks. esimerkiksi Alvesalo-Kuusi, A., & Tolsa, T. (2022). Mihin kriminologin katse kohdistuu – Vastaamo-tapauksen monet kasvot. (24.11.2022). *Haaste*.

20 Ks. esimerkiksi Fiorentino, V., Harrikari, T., Saraniemi, S., & Romakkaniemi, M. (2023). Sosiaalialan työn kiihtyvyys COVID-19–pandemian seurauksena. *Sosiaalityö kriiseissä.* 243.

Tekoälyyn liittyvä sanasto on suomeksi laajaa, mutta jokseenkin vakiintumatonta. Jotta lukijalla olisi mahdollisuus etsiä lisää tietoa käsittelemistäni teemoista, ja jotta tutkimusta voisi arvioida myös kansainvälisessä kontekstissa, tuon esille vakiintuneet englanninkieliset termit suomenkielisten termivalintojeni yhteydessä.

1.1 Algoritmiset teknologiat yhteiskunnallisten haittojen tutkimuksessa

Yhteiskunnallisten haittojen teoria pyrkii laajentamaan ymmärrystä haitoista yksilö- ja tekokeskeisen rikosoikeusjärjestelmän ja sitä tutkivan valtavirtakriminologian ulkopuolelle. Yhteiskunnalliset haitat viittaavat siis haittoihin, jotka syntyvät yhteiskunnan erilaisten prosessien seurauksena, mutta joita rikosoikeusjärjestelmän puitteissa ei tunnisteta tai säännellä. Teoria soveltuu tekoälyjärjestelmien aiheuttamien haitallisten vaikutusten tutkimiseen erinomaisesti. Algoritmisten teknologioiden muutosvoima on niin valtava, että puhutaan neljännestä teollisesta vallankumouksesta[21]. Teknologiat muokkaavat radikaalilla tavalla niin aikaa ja tilaa kuin politiikkaa, etiikkaa ja ympäristöäkin, samoin kuin ihmisen toimintaa, ymmärrystä ja identiteettiä. On jokseenkin selvää, että muutokset voivat synnyttää paitsi hyötyjä myös haittoja, joita rikosoikeusjärjestelmä ei ainakaan toistaiseksi tunnista tai sääntele.

Yhteiskunnallisten haittojen taustalla vaikuttavat usein kapitalistiselle järjestelmälle ominaiset seikat: vallan ja varallisuuden epätasainen jakautuminen, kilpailu ja elämän eri osa-alueiden kaupallistaminen. Kriittisen kriminologian ja etenkin yhteiskunnallisten haittojen tutkimuksen parissa näiden katsotaan vaikuttavan haitallisesti ihmisiin ja ryhmiin, riistävän hyvinvoinnin edellytyksiä ja altistavan laajasti erilaisille haitoille[22]. Sen sijaan valtavirtakriminologian piirissä tällaiset haitat nähdään usein luonnollisina, väistämättöminä ja tahattomina[23]. Etenkin yritysten ja valtion toiminnasta johtuvat, laajalle levittäytyvät ja hitaasti kehittyvät haitat ovat perinteisesti jääneet kriminologian tutkimuskysymysten ulkopuolelle[24].

Yhteiskunnallisten haittojen tutkimuksen pyrkimyksenä on yhteiskunnallisten, haitallisten prosessien kokonaisvaltainen muutos, jonka saavuttamiseksi epätasaiset

21 Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

22 Whyte, D. (2017). Crime as a social relation of power: Reframing the 'ideal victim'of corporate crimes. In *Handbook of Victims and Victimology* (pp. 333–347). Routledge.

23 Pemberton, S. A. (2015). *Harmful societies: Understanding social harm*. Policy Press. 3–5;
Friedrichs, D. O. (2009). Trusted criminals: White collar crime in contemporary society. Cengage Learning.

24 Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press. 19.

valtarakenteet ja niiden haitalliset vaikutukset pyritään tuomaan näkyville. Kuten kriittisissä tutkimusperinteissä usein, yhteiskunnallisten haittojen tutkijan kiinnostus suuntautuu usein yhteiskunnassa esiintyviin toimintoihin, jotka toisintavat, lisäävät ja luovat keinoja kerryttää valtaa ja varallisuutta muiden kustannuksella. Tällaisten toimintojen ja niistä seuraavien haittojen hallitsemisessa myös lainsäädännöllä on merkitystä. Tämän takia haittojen tutkimus kiinnittyy myös oikeustieteeseen: sen avulla voidaan paljastaa lainsäädännöstä seikkoja, jotka mahdollistavat haitallisia toiminnan tapoja tai kannustavat niihin. Haitallisiin toimiin kuuluvat luonnollisesti myös toimimattomuus ja toiminnan laiminlyönti silloin, kun ne synnyttävät tai ylläpitävät haittoja.

Algoritmisen, nopean ja suurelta osin hallitsemattoman transformaation ja siihen kiinnittyvien yhteiskunnallisten muutosten käsittelyyn normatiivinen oikeustiede tarjoaa vain rajallisesti työkaluja. Etenkin haittoja aiheuttavat muutokset, joiden taustalta ei voida erottaa kriminalisoituja tekoja tai laiminlyöntejä, jäävät usein oikeudellisen ja myös kriminologisen tutkimuksen ulkopuolelle. Haitan käsitteen kautta epätoivottuja muutoksia ja niiden seurauksia voidaan arvioida säädännäisestä oikeudesta riippumatta. Tällöin on mahdollista esimerkiksi tutkia joustavasti uusia ilmiöitä, joita ei vielä ole kattavasti säädelty, tai tunnistaa olemassa olevan sääntelyn heikkoja kohtia tai puutteita, jotka altistavat haittojen ilmenemiselle.

Toisaalta on väistämätöntä, että mikäli haittoja käsitellään oikeudellisessa kontekstissa, mutta erillään rikoksen kriminalisointeihin sidotusta konseptista, esiin nousee kysymys siitä, miten haitta määritellään[25]. Samalla kun haitan määritelmän joustavuus, joka toisinaan lähenee myös epäselvyyttä, mahdollistaa haitallisiksi arvioitujen, mutta rikoksiksi määrittelemättömien tai heikosti säänneltyjen ilmiöiden tunnistamisen ja ennakkoluulottoman tutkimisen, se voi herättää kysymyksiä myös yhteiskunnallisten haittojen tutkimuksen uskottavuudesta[26]. Jos haitan määritelmä sidotaan subjektiiviseen haitan kokemukseen, se törmää näkemysten ja kokemusten ristiriitaisuuteen ja tulkinnallisuuteen. Jos tämä halutaan välttää ja haitalle määritellään rajattu kriteeristö, joustavuus kärsii ja mahdollisuudet haitallisten ilmiöiden laajaan tunnistamiseen ja tutkimiseen rajautuvat samoin kuin kiinnityttäessä rikosten

25 Ks. esimerkiksi Lasslett, K. (2010). Crime or social harm? A dialectical perspective. *Crime, Law and Social Change*, *54*, 1–19;

Yar, M. (2012). Critical criminology, critical theory and social harm. Teoksessa Hall, S. & Winlow, S. (Eds). *New directions in criminological theory*. Routledge. 52–65;

Raymen, T. (2022). The enigma of social harm: The problem of liberalism. Taylor & Francis.

26 Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press. 20.

määritelmiin; seikka, jota yhteiskunnallisten haittojen tutkimuksen parissa on arvosteltu laajasti[27].

Joissakin yhteyksissä yhteiskunnallisten haittojen tutkimusta on kritisoitu liiallisesta kiinnittymisestä liberaaliin ideologiaan ja subjektiivisiin[28], jopa emotivistisiin käsityksiin haitasta[29]. Kritiikissä esitetään, että uskottavuuden ja yleistettävyyden takaamiseksi tulisi määritellä haitan todellinen olemus ja ontologia[30], jotta haittojen tutkimus ei perustu tutkijan mielivaltaiseen päätökseen siitä, mitä haitta milloinkin on, kenen kannalta asiaa arvioidaan ja kenen näkökannat otetaan huomioon. Kritiikki on sinänsä perusteltua. Yhteiskunnallisten haittojen tutkimuksessa haitan ontologiaa ei ole määritelty tiukasti. Erilaisia viitekehyksiä on useita[31], ja haitallisuutta on arvioitu niin (positiivisten) oikeuksien, inhimillisten tarpeiden kuin moraalisten velvoitteiden kautta. Monissa kriittisissä tutkimussuuntauksissa, kuten feministisessä, postkolonialistisessa ja queertutkimuksessa, hyödynnetään yhteiskunnallisten haittojen teorian käsitteistöä. Luonnollisesti käsitteistön merkityssisältö on siis melkoisen vaihtelevaa.

Yhteiskunnallisten haittojen tutkimuskentällä kenties laajimmin hyödynnetyssä määritelmässä haitallisuus määräytyy suhteessa inhimillisiin tarpeisiin[32]. Näitä tarpeita loukkaavat esimerkiksi haitat, jotka kohdistuvat fyysiseen tai psyykkiseen hy-

27 Hillyard ja Tombs (2004) argumentoivat, että rikokselta puuttuu ontologinen todellisuus, ja näin ollen valtavirtakriminologian tutkijat sitoutuvat sosiaalisesti luotuun konstruktioon, ie. mielikuvitukseen. Ks. ibid. s. 11.

28 Hall, S., & Winlow, S. (2018). Ultra-realism. Teoksessa *Routledge Handbook of Critical Criminology* (2. painos). 43–56. Routledge.

29 Raymen, T. (2022). The enigma of social harm: The problem of liberalism. Taylor & Francis. 7.

30 Hall, S., & Winlow, S. (2018). Ultra-realism. Teoksessa *Routledge Handbook of Critical Criminology* (2. painos). 43–56. Routledge;
Raymen, T. (2022). The enigma of social harm: The problem of liberalism. Taylor & Francis. 7.

31 Ks. esimerkiksi Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press;
Yar, M. (2012). Critical criminology, critical theory and social harm. Teoksessa Hall, S. & Winlow, S. (eds). *New directions in criminological theory*. Routledge;
Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

32 Yar, M. (2012). Critical criminology, critical theory and social harm. Teoksessa Hall, S. & Winlow, S. (eds). *New directions in criminological theory*. Routledge;
Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

vinvointiin, autonomiaan tai suhteelliseen asemaan (*relational harms*)[33]. Toisaalta voidaan argumentoida, että haittojen erilaiset kategoriat eivät vielä riitä määrittämään haitan perustavanlaatuisia ominaisuuksia. Esimerkiksi Raymen[34] katsoo, että haitan määrittelemiseksi tarvitsemme väistämättä yhteisen käsityksen siitä, mikä on *hyvää* tai toivottavaa. Tämän määrittelemisessä hän soveltaa (neo)aristoteelistä hyve-etiikkaa, johon tukeutuen hän määrittelee haitan ontologian hyveen negaationa: mikäli toimintaa ei voi rationaalisesti perustella hyväksi sen itsensä vuoksi, se voidaan arvioida haitalliseksi riippumatta yksilön henkilökohtaisesta suhtautumisesta. Määritelmässä etäännytään yksilön kokemuksesta ja haetaan vakuuttavuutta filosofian perinteistä.

Hyve-etiikan perinteeseen tukeutuvassa määritelmässä on potentiaalia, mutta siinä on myös riskinsä. Jos hyvä tai hyve määritellään ylhäältä käsin, voidaan päätyä arvioimaan jokseenkin elitistisesti, milloin kokemus haitasta tai koettu kärsimys on todellista ja huomion arvoista. Perinteisiä aristoteelisia hyveitä eli käytännöllistä viisautta, rohkeutta, oikeamielisyyttä ja kohtuullisuutta ei monituhatvuotisessa perinteessä ole onnistuttu yksiselitteisesti määrittelemään. Haitan käsitteen sitominen aristoteeliseen hyveen määritelmään törmää siis subjektiivisiin näkemyksiin samoin kuin sen sitominen suoraan haitan ominaisuuksien määrittelyyn. Vahva tukeutuminen hyve-etiikkaan voi myös johtaa siihen, että haitallisuus määritellään samalla moraalisesti paheksuttavaksi. Toisaalta tällainen määritelmä antaa mahdollisuuden lähestyä myös ilmiöitä, joita ei välttämättä koeta haitallisiksi, mutta jotka voivat siitä huolimatta luoda ongelmia tai vaivihkaa heikentää mahdollisuuksia kukoistavaan elämään, kuten esimerkiksi monia harmittomina ja hyödyllisenä pidettyjä uusia teknologioita.

Haitan määrittelyn joustavuus on toisaalta myös yhteiskunnallisten haittojen tutkimuksen vahvuus[35]. Joustavan haittakäsityksen ansiosta on mahdollista huomioida valta-asetelmat ja nostaa esiin yhteiskunnallisia prosesseja, jotka vaikuttavat esimerkiksi eriarvoisuuden ja siihen kiinnittyvien haittojen taustalla, ja jotka heijastavat

33 Pemberton, S. A. (2015). *Harmful societies: Understanding social harm*. Policy Press. Ks. myös Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press. Hillyard ja Tombs jaottelevat haitat fyysisiin, taloudellisiin ja kulttuurisiin.

34 Raymen, T. (2022). *The enigma of social harm: The problem of liberalism*. Taylor & Francis. 233. Määritelmä kiinnittyy vahvasti rationaalisuuteen, joka taas kiinnittyy globaalin pohjoisen filosofisiin perinteisiin. Näennäinen objektiivisuus ei siis nähdäkseni poista tästäkään määritelmästä ideologisia perusteita, vaan sitoumukset ovat ainoastaan muualla kuin yksilökeskeisessä (arvo)liberalismissa.

35 Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press. 20.

vaikutuksensa myös säädännäiseen oikeuteen. Teorian parissa on tutkittu laajasti juuri vähemmistöjen, marginalisoitujen, syrjäytyneiden tai muutoin yhteiskunnan vähäosaisten kohtaamia haittoja, jotka syntyvät vallan epätasaisen jakautumisen seurauksena. Tämä tietenkin tekee yhteiskunnallisten haittojen tutkimuksesta osaltaan poliittista[36]. Voidaan nähdä, että tavoitteena on nimenomaan tunnistaa haitallisia ilmiöitä ja arvioida yhtäältä, mistä sosiaalisista tai yhteiskunnallisista prosesseista haitat kumpuavat ja toisaalta, mihin ne johtavat tai voivat johtaa. Yhteiskunnallisten haittojen tutkimuksen avulla on siis mahdollista argumentoida, milloin ja miksi yhteiskunnallinen muutos olisi tarpeellinen tai välttämätön.

Tässä työssä haitat määritellään suhteessa ihmisen mahdollisuuksiin toteuttaa hyvää elämää. Hyvä elämä taas määräytyy inhimillisten tarpeiden kautta; ei niiden maksimaalisen täyttymisen vaan huonommillaankin *riittävän* täyttymisen kautta. Näkökulmassa on nähtävissä hyve-etiikan vaikutus, sillä subjektiivisen kokemuksen lisäksi tärkeäksi nousee pyrkimys jakaa vähintään riittävästi hyvinvointia kaikille: perinteiset aristoteeliset hyveet etenkin oikeamielisyydestä ja kohtuudesta vaikuttavat siihen, mitä voidaan pitää jokaiselle tarpeellisina hyvän elämän edellytyksinä.

Kun arvioin hyvän elämän mahdollisuuksia, nyky-yhteiskunnassa olennaisiksi seikoiksi nousevat erityisesti 1) taloudellinen turvallisuus, 2) fyysinen ja psyykkinen terveys ja hyvinvointi, 3) autonomia ja 4) mahdollisuudet yhteiskunnassa ja suhteessa muihin. Kategorioissa yhdistyvät Pembertonin[37] sekä Hillyardin ja Tombsin[38] haittojen typologiat[39]. Olennaiset seikat on pyritty tiivistämään siten, että ne sisältävät hyvän elämän perusedellytykset.

Hyvän elämän edellytyksiä on tutkittu modernissa filosofiassa paljon. Esimerkiksi rawlsilaista oikeudenmukaisuusteoriaa ja hyve-etiikkaa yhdistelevä Nuss-

36 Tämä lienee osaltaan syy sille, että esimerkiksi ultarealistit kritisoivat tutkimusalaa väittäen, että se perustuu emotivismiin. Ks. Ilan, J. (2019). Cultural criminology: The time is now. *Critical Criminology*, *27*, 5–20. Ilan huomauttaa, että kun toiset argumentoivat, että esimerkiksi sukupuoli ja etninen tausta vaikuttavat *valtasuhteisiin*, etenkin ultrarealistit argumentoivat, että tällainen argumentaatio lähestyy identiteettipolitiikkaa.

37 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

38 Tombs, S. ja Hillyard, P. (2004). Towards a political economy of harm: States, corporations and the production of inequality. Teoksessa Hillyard, P., Pantazis, C., Tombs, S. & Gordon, D. (eds). *Beyond Criminology: Taking harm seriously*: 30–54. London: Pluto Press.

39 Tombs ja Hillyard (ibid.) luokittelevat haitat seuraavasti: 1) fyysiset haitat, 2) taloudelliset ja varallisuuteen kohdistuvat haitat, 3) psyykkiset haitat ja 4) haitat kulttuuriselle turvallisuudelle; Pemberton (ibid.) katsoo haittojen jakautuvan seuraavasti: 1) fyysiseen ja psyykkiseen terveyteen kohdistuvat haitat, 2) autonomiaan kohdistuvat haitat ja 3) haitat suhteessa muihin ja ympäristöön (relational harms).

baum[40] hakee kukoistavan elämän perusteita toimintamahdollisuuksien (*capabilities*) kautta. Nussbaumin mukaan ihmisellä tulee olla mahdollisuus (1) elämään, (2) fyysiseen terveyteen, (3) integriteettiin, (4) aisteihin, mielikuvitukseen ja ajatteluun, (5) tunteisiin, (6) järkeilyyn, (7) yhteyteen muihin, (8) muiden lajien ja ympäristön huomioimiseen, (9) leikkiin ja (10) oman ympäristönsä hallintaan niin poliittisessa kuin materiaalisessa mielessä. Nämä toimintamahdollisuudet on mahdollista sijoittaa esittämääni nelikohtaiseen haittojen typologiaan. Listan ensimmäiset kuusi kohtaa kiinnittyvät etenkin psyykkiseen ja fyysiseen terveyteen ja hyvinvointiin sekä autonomiaan, kohdat 7 ja 8 mahdollisuuksiin toimia yhteiskunnassa ja suhteessa muihin, kohdat 9 ja 10 taas autonomiaan. Kohta 10 liittyy osaltaan myös taloudelliseen turvallisuuteen. Toisaalta taloudellinen turvallisuus on kapitalistisessa maailmassa välttämätön edellytys lähes kaikkien toimintamahdollisuuksien toteuttamiseksi.

Yhteiskunnallisten haittojen tutkimuksessa yhteiskuntaa prosesseineen on harvoin arvioitu muuten kuin nimenomaan haittoja luovana entiteettinä, jonka toimintalogiikka on syy haittojen ilmenemiselle[41]. Laajennan näkökulmaa huomioimalla, että yhteiskunta voi olla myös *haittojen kohde.* Ideaalitilanteessa yhteiskunnan tulisi luoda edellytykset sille, että kansalaisten olisi mahdollista elää mahdollisimman hyvää, tarpeidensa mukaista elämää[42]. Käsittelenkin tässä johdanto-osassa myös yhteiskuntaan kohdistuvia, tekoälyteknologioihin tai niiden käyttöön liittyviä haittoja (*societal harm*[43]), joilla on potentiaalia muuttaa yhteiskuntaa tai sen toimintamahdollisuuksia siten, että yhteiskunnan edellytykset kansalaisten hyvän elämän turvaamiseen heikkenevät.

Siinä missä yksilöön kohdistuvat haitat vaikuttavat välittömästi yksilön hyvän elämän edellytyksiin, yhteiskuntaan kohdistuvat haitat voivat horjuttaa yhteiskunnan vakautta ylläpitäviä rakenteita ja järjestelmiä, kuten demokraattisia valtarakenteita tai talous- tai oikeusjärjestelmiä. Muutokset voivat haastaa sosiaalisia ja kulttuurisia normeja tavoilla, jotka muokkaavat laaja-alaisesti yhteiskunnan toimintaa ja sitä myötä myös ihmisten mahdollisuuksia. Kaikki muutokset eivät kuitenkaan automaattisesti ole haitallisia, päinvastoin. Modernin läntisen maailman kapitalistiset

40 Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

41 Esimerkiksi Raymen katsoo, että tämä on seurausta siitä, että haittaa ei ole onnistuttu määrittelemään onnistuneesti. Ks. Raymen, T. (2022). *The enigma of social harm: The problem of liberalism*. Taylor & Francis. 223.

42 Esimerkiksi Nussbaum on argumentoinut näin, ks. Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

43 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).

valtiot eivät tälläkään hetkellä tarjoa kattavia edellytyksiä kansalaistensa hyvälle, kukoistavalle elämälle, kuten monet tutkijat ovat vakuuttavasti argumentoineet[44]. Muutos voi siis monissa tilanteissa olla toivottu ja haittoja vähentävä tekijä, jopa tavoite, kuten aiemmin on esitetty. Kuitenkin muutokset, joiden seurauksena yhteiskunta pystyy yhä heikommin turvaamaan kansalaisilleen hyvän elämän edellytykset, tulee tunnistaa haitallisiksi ja ottaa vakavasti myös yhteiskunnallisten haittojen tutkimuksen parissa.

Algoritmisilla teknologioilla on väistämättä vaikutuksia erilaisten yhteiskunnallisten haittojen syntymiseen ja haittojen ominaispiirteisiin, kuten ensimmäisessä osatutkimuksessa esitetään. Algoritmisten järjestelmien tekniset ominaisuudet, joita Viljanen[45] on analysoinut, määrittävät osaltaan välittömässä käyttöympäristössä ilmenevien haittojen muodostumista. Algoritmiset teknologiat muovaavat yhteiskunnan sosioekonoteknisiä järjestelmiä kuitenkin välitöntä käyttöympäristöään laajemmin, ja on jo nähtävillä, että muutos ei tapahdu ainoastaan kohti parempaa. Muutosten seurauksena mahdollisesti syntyvät haitat ovat kuitenkin vielä jääneet vaille yhteiskunnallisten haittojen tutkijoiden suurempaa huomiota, ja monien algoritmisten haittojen ominaispiirteet ja vaikutukset sekä niiden syntyyn vaikuttavat seikat ovat ainakin osittain tunnistamatta.

1.2 Tekoäly muutoksen kiihdyttäjänä

Hartmut Rosan mukaan modernit yhteiskunnat tunnistaa siitä, että niitä säätelee, koordinoi ja hallitsee tiukka ajallinen järjestelmä[46], jonka keskiössä eivät ole oikeudenmukaisuuden tai etiikan säännöstöt. Vaikka tällainen yhteiskunta on yhtäältä mahdollisimman vapaa ja sen toimintoja rajoittavat äärimmäisen vähän eettiset periaatteet tai normistot, toisaalta sen koko olemassaolo on alisteinen näkymättömälle, depolitisoituidulle, keskustelemattomalle, aliteoretisoidulle ja artikuloitumattomalle aikajärjestelmälle, joka määrittää reunaehdot sen toiminnalle[47]. Tätä aikajärjestelmää Rosa analysoi yhteiskunnallisen kiihtyvyyden käsitteen avulla.

Rosan näkemyksen mukaan kiihtyvyys toimii näkymättömänä voimana yhteiskunnan muutosten taustalla. Näkymättömyytensä vuoksi se myös vaikuttaa neutraalilta: sitä on vaikea määritellä suhteessa hyvään tai toivottavaan. Se näyttäisi perus-

44 Esimerkiksi Pemberton, S. A. (2015). *Harmful societies: Understanding social harm*. Policy Press.

45 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

46 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

47 Rosa, H. (2010). Alienation and Acceleration. Towards a Critical Theory of Late-Modern Temporality. NSU Press. 8.

tuvan jatkuvan kehityksen ideaalille, sillä onhan selvää, että yhteiskunnan toimintojen ja myös inhimillisen elämän kehitys kohti parempaa vaatii uusia tieteellisiä ja teknologisia innovaatioita, joiden avulla yhä suurempi osa maailmasta ja sen mahdollisuuksista pystytään tuomaan jokaisen ulottuville. Kiihtyvyyden logiikka kuitenkin toimii tästä ideaalista erillään. Sen perustavanlaatuinen merkitys tulee esille, kun arvioidaan, mitä moderni yhteiskunta vaatii pysyäkseen vakaana ja pystyäkseen suoriutumaan tehtävistään. Rosa puhuu dynaamisesta vakauttamisesta (*dynamic stabilization*) modernin yhteiskunnan perustavanlaatuisena tuntomerkkinä: ei ainoastaan kehittyäkseen, vaan myös ylläpitääkseen olemassa olevaa yhteiskunnan eri prosessien täytyy toimia jatkuvasti nopeammin, tuottaa koko ajan enemmän ja luoda yhä uusia innovaatioita ja toimintamalleja. Tämä väistämätön vaatimus koskettaa yhteiskunnan kaikkia sektoreita:

> Without expansion, innovation and accumulation, companies close down, jobs are lost, and, by consequence, public revenues decrease and expenditures increase, and the ensuing monetary and fiscal crisis can put political legitimation at risk, too. […]
>
> It is important here to note that 'dynamic stabilization' in this sense involves more than just permanent processing or continuous operation. It is intrinsically tied to the logic of increase, i.e. stability – albeit an oftentimes shaky and rather temporary form of stability – derives from growth, augmentation and efficient innovation, not just from processual reproduction.
>
> It might be debatable whether the capitalist engine is the only motor for this mode of dynamic stabilization. In fact, it can be observed in the way science, politics, and art are conceived, conducted and reproduced in modern societies, too. Thus, modern science is no longer about the preservation and tradition of (ancient, holy, etc.) knowledge, but about perpetual 'progress' toward new questions, projects and answers, while modern art is bent on innovation and transgression instead of imitation or approximation. Similarly, legislation has become a perennial task of innovative adaption, and democratic governmental rule by definition requires dynamic renewal and confirmation to keep the system stable.[48]

[48] Rosa, H., Dörre, K., & Lessenich, S. (2017). Appropriation, activation and acceleration: The escalatory logics of capitalist modernity and the crises of dynamic stabilization. *Theory, Culture & Society*, *34*(1), 53–73.

Kiihdyttäminen on siis perustavanlaatuinen vaatimus modernin yhteiskunnan säilyttämiseksi. Se kuvaa ajan suhdetta muutokseen, ja vaatimuksena on, että aika per tehtävä lyhenee jatkuvasti eli todellisuus *tiivistyy ajassa*. Myös algoritmiseen transformaatioon liittyy väistämättä vahvasti ajallinen elementti – transformaatio tapahtuu *ajan kuluessa*, ja muutos on kiihtynyt niin ajassa kuin tilassa arvioiden digitalisaation, algoritmisaation ja viimeisimpänä tekoälyn kehityksen ja käyttöönoton myötä. Tekoälyteknologiat ovat teknologisten innovaatioiden eli kiihtyvän teknologisen kehityksen viimeisimpiä saavutuksia. Niillä on potentiaalia vaikuttaa sekä siihen, *millä tavalla* yhteiskunta muuttuu, että siihen, *kuinka nopeasti* muutos tapahtuu. Uudet ja yhä kehittyvät teknologiat tarjoavat loputtomia mahdollisuuksia uudistaa, nopeuttaa ja tehostaa toimintoja niin valtioiden hallinnossa, yksityisellä sektorilla kuin myös ihmisten jokapäiväisessä elämässä, töissä ja vapaa-ajalla.

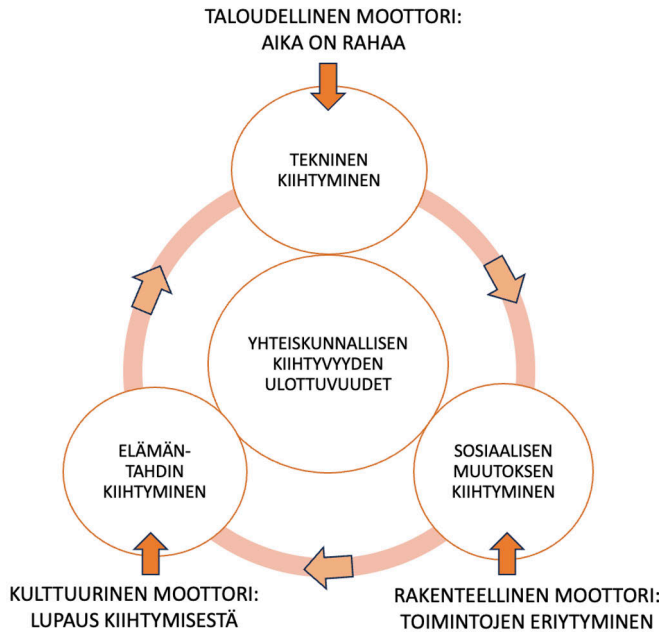
Ajassa vähitellen tapahtuvan muutoksen hallinta lainsäädännön keinoin on lähtökohtaisesti hankalaa. Tätä hankaluutta ei vähennä se, että muutoksia tapahtuu monilla eri tasoilla ja eri tavoilla, ja muutokset vaikuttavat toisiinsa. Perinteisesti sääntelyllä pystytään vaikuttamaan asioihin *ex ante* tai *ex post*: sääntelyllä voidaan pyrkiä estämään epätoivottuja käytänteitä ennakolta tai puuttumaan niihin jälkikäteisen kontrollin keinoin. Itse muutoksen hallitseminen sen sijaan jää usein sääntelyn ulottumattomiin[49]. Tämä on nähty historian aikana lukemattomia kertoja, kun uudet innovaatiot ovat pyyhkäisseet yli yhteiskuntien ja muovanneet toimintoja uusiin suuntiin[50]. Epätoivottujen muutosten hallitsemiseksi tai edes hillitsemiseksi on kuitenkin äärimmäisen tärkeää pyrkiä ymmärtämään muutosta ja ennakoimaan siitä kumpuavia haitallisia seurauksia, jotta sääntelyä pystytään kohdentamaan merkityksellisiin seikkoihin.

Algoritmista transformaatiota voidaan arvioida paitsi yhteiskunnallisen kiihtyvyyden logiikasta voimansa saavana monipuolisena ja vaikeasti ennakoitavissa olevana kehityskulkuna, myös kiihtymistä yhä voimistavana ilmiönä. Kiihtyvyys rakentuu kehämäiseksi tapahtumaketjuksi (kuva 1), jossa teknologinen kehitys on paitsi seurausta kiihtyvyyden logiikasta, myös merkittävä syy kiihtyvyyden voimistumiselle. Tämän tuplaposition takia teknologiset innovaatiot määrittävät yhteiskunnan muutosta keskeisellä, vaikkakaan eivät sinänsä ennenkuulumattomalla tavalla[51]. Alla oleva kuva selventää kiihtyvyyden moottorien välisiä vuorovaikutussuhteita.

49 Hasselbalch, J. A. (2018). Innovation assessment: governing through periods of disruptive technological change. *Journal of European Public Policy*, *25*(12), 1855–1873.

50 Ks. esimerkiksi Moses, L. B. (2007). Recurring dilemmas: The law's race to keep up with technological change. *U. Ill. JL Tech. & Pol'y*, 239.

51 Myös STS:n parissa vuorovaikutussuhde on havaittu ja sitä on käsitteellistetty *co-production*-termin avulla: yhteiskunnallinen todellisuus sekä tieto ja sen sovellukset, kuten modernit teknologiat, määrittävät toisiaan ja kehittyvät vuorovaikutussuhteessa toisiinsa. Ks. Jasanoff, S. (2004). The idiom of co-production. Teoksessa *States of knowledge* (1–12). Routledge.



Kuva 1. Kiihtyvyyden moottorit yhteiskunnallisen kiihtyvyyden teorian mukaan[52].

Rosan mukaan kiihtyvyyttä ruokkiva tapahtumaketju syntyy kolmen niin kutsutun moottorin ja niiden luomien vaikutusten kautta: kapitalistinen talousjärjestelmä toimii taloudellisena moottorina, joka kiihdyttää teknologioiden (laajasti ymmärrettynä) kehittämistä. Toimintojen eriytyminen (*functional differentiation*) voidaan ymmärtää rakenteelliseksi (*sociostructural*) moottoriksi, joka kiihdyttää muutosta yhteiskunnassa: kun eri tehtävät ja eri toiminnot irtoavat toisistaan, monimutkaisuus lisääntyy, mutta mahdollisuudet erikoistumiseen ja sen tuomiin etuihin, kuten innovaatioiden lisääntymiseen ja tuotannon tehostamiseen paranevat. Lupaus tehokkaammista ja *paremmista* ratkaisuista taas luo kulttuurisen moottorin, joka kannustaa ihmisiä kiihdyttämään elämäntahtia jokapäiväisessä elämässään. Nämä erilliset prosessit ruokkivat toisiaan, ja yhden osa-alueen kiihdyttäminen vaatii muita osa-alueita reagoimaan.[53] Esimerkkejä on lukuisia. Toimintojen eriytyminen tehostaa tuotantoa, tehostunut tuotanto ruokkii taloudellista moottoria ja siten tuo uusia teknologioita nopeammin ja laajemmin saataville, ja nopeammat teknologiat mahdollistavat arkisen elämän ajallisen tiivistymisen. Esimerkiksi auto mahdollistaa nopeammat siirtymiset, jolloin autoileva ihminen pystyy liikkumaan pidemmälle tai siirtymään paikasta toiseen nopeammin. Tällöin vuorokauteen mahtuu useampia

52 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 194.
53 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

erillisiä tapahtumia. Mahdollisuus hoitaa asiat tietokoneen välityksellä tuottaa saman vaikutuksen, mutta vielä voimakkaampana: esimerkiksi etätyöläinen pystyy osallistumaan vaikkapa useampaan kokoukseen samanaikaisesti. Kiihtyvä vauhti antaa ideaalitilanteessa mahdollisuuksia paitsi miellyttävämpään ja sujuvampään arkeen myös tuottavampaan toimintaan. Toisaalta se lisää riskiä sille, että aika niin sanotusti valuu jatkuvasti kiihtyvällä tahdilla läpi sormien: vaikka innovaatiot lisäävät mahdollisuuksien ja tosiasiallisen tekemisen määrää, samalla lisääntyy kokemus siitä, että yhä enemmän merkittäviä asioita jää tekemättä ja kokematta[54]. Jatkuvasti kaventuva aikaikkuna eli vaatimus yhä välittömämmästä reagoinnista ja tehokkaammasta toiminnasta heikentää niin yksilöiden kuin instituutioiden mahdollisuuksia arvioida ja toteuttaa merkityksellisiä, oikeudenmukaisia ja eettisesti perusteltuja ratkaisuja.

Rosan kuvaama kehämäinen logiikka, joka ruokkii kiihtymistä, nostaa näkyville yhteiskunnassa vallitsevat monisuuntaiset vuorovaikutussuhteet. Samoin kuin Rosan kiihtyvyysteoriassa, myös Giddensin rakenteistumisteoriassa erisuuntaiset vuorovaikutussuhteet korostuvat. Rakenteistumisteoriassa argumentoidaan, että yksilön ja yhteiskunnan välinen suhde on duaalinen eli kaksinainen, ei dualistinen: yksilö ja yhteiskunta toimivat ja muovautuvat vuorovaikutuksessa toisiinsa ja määrittävät toisiaan. Siinä missä yhteiskunta vaikuttaa toimijaan eli yksilöön, myös toimija vaikuttaa yhteiskuntaan, ja molemmat vaikutussuunnat ovat yhtä merkityksellisiä.[55] Samankaltainen lähestymistapa näkyy kiihtyvyysteoriassa, jossa Rosa painottaa yksilöiden aktiivista roolia kiihtyvyyden logiikan ylläpitäjinä sen sijaan, että näkisi heidät ainoastaan yhteiskuntajärjestyksen uhreina[56]. Giddensin mukaan yhteiskunnassa vallitsee tietty järjestys, ajasta ja paikasta riippumaton *rakenne (structure),* joka vaikuttaa yhteiskunnallisten järjestelmien (*social systems*) taustalla. Järjestelmät ilmentävät rakennetta ja mahdollistavat sosiaalisten käytänteiden syntymisen. Toimija toimii rakenteen sallimissa rajoissa, mutta samalla muovaa tarkoituksella ja tahattomasti yhteiskunnallisia käytänteitä ja siten myös järjestelmää ja lopulta rakennetta vahvistaen, heikentäen tai muuttaen yhteiskunnassa vallitsevia, osin tiedostamattomia toiminnan ehtoja. Näin ollen toiminnat muodostuvat osaksi institutionaalisia käytänteitä – ja nämä käytänteet

54 Ibid. 136.

55 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

56 Rosa katsoo, että ihmisten toimintaa motivoi pyrkimys hyvään elämään, ja hyvän elämän määrittävinä ominaisuuksina on ”kolme A:ta”: se on ”available, accessible, attainable” eli saatavilla, saavutettavissa ja osallistumismahdollisuuksien rajoissa. Pyrkimys kohden aina vain saavutettavampia mahdollisuuksia ruokkii yhteiskunnan kiihtymistä. Ks. Rosa, H. (2017). Available, accessible, attainable: The mindset of growth and the resonance conception of the good life. Teoksessa *The Good life beyond growth* (39–53). Routledge.

taas ylläpitävät rakenteita, joiden puitteissa toiminta yhteiskunnassa tapahtuu. Olennaista on ymmärtää, että muutos voi olla paitsi tarkoituksellinen myös tahaton ja tiedostamaton – jolloin sen suunta ei välttämättä ole ennalta arvattavissa. Duaalista muutosprosessia Giddens kutsuu rakenteistumiseksi (*structuration*).[57]

Yhteiskunnallisen kiihtyvyyden teoria vaikuttaa sopivan rakenteistumisteorian kanssa yhteen, ja kiihtyvyyden hallitsemisen monipolviset haasteet on helppo ymmärtää, kun teorioita tarkastelee rinnakkain. Kiihdyttämistä, tehostamisesta ja kehitystä ajavat sekä tietoiset että tiedostamattomat valinnat, joita ihmiset ja instituutiot tekevät yhteiskunnan mahdollistamissa puitteissa, ja joiden kautta he samalla vaikuttavat yhteiskunnallisen muutoksen suuntaan ja nopeuteen. Tällainen duaalinen vuorovaikutus on rakenteistumisteorian ytimessä. Kaikkia seikkoja, jotka yhtäältä luovat puitteet kiihdyttämiselle ja toisaalta edistävät sitä yksittäinen toimi kerrallaan, on äärimmäisen vaikea hallita. Tarve kiihtyvän vauhdin hillitsemiselle kuitenkin kasvaa. Yhteiskunnan eri sektoreilla saman muutosvauhdin – ja myös kiihtymistahdin – ylläpitäminen käy koko ajan vaikeammaksi. Eri toimintojen kiihtyvyyttä rajoittavat erilaiset tekijät eri tavoin, ja yhä useammassa tapauksessa keinovalikoima kiihtyvän vauhdin ylläpitämiseksi loppuu kesken. Esimerkiksi valtiollista päätöksentekoa on äärimmäisen vaikea toteuttaa enää nykyistä nopeammin tinkimättä demokraattisista periaatteista. Toisaalta taas esimerkiksi teknologinen kehitys on etenkin viime vuosina ollut äärimmäisen nopeaa, eivätkä kiihtyvyyden rajat eivät vielä näyttäisi olevan tulossa vastaan. Toisaalta teknologiseen kehitykseen tarvittavat raaka-aineet hupenevat monien muiden luonnonvarojen tavoin tai niiden saatavuus esimerkiksi ympäristösääntelyn tiukentuessa heikkenee, kuten olemme nähneet esimerkiksi erikoisempien metallien kohdalla[58]. Tästä huolimatta teknologinen kehitys on niin nopeaa, että demokraattisissa prosesseissa väistämättä verkkaisesti muutettavissa olevan lainsäädännön mahdollisuudet hallita näitä kehityskulkuja ovat heikentyneet. Tämä ilmiö näyttäytyy erityisen selkeänä tekoälyn sääntelyn pyrkimyksissä: riskinä on, että lainsäädäntö demokraattisten prosessien jälkeen voimaan tultuaan on hyvin nopeasti vanhentunutta. Tällöin lainsäädäntö ei kykene toimimaan kiihtymisvauhdin vakauttajana (*stabilizer*) tai hidastajana (*decelerator*), vaan muutoksen hallinta karkaa sen ulottumattomiin[59].

57 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

58 Ks. esimerkiksi Zeng, A., Chen, W., Rasmussen, K. D., Zhu, X., Lundhaug, M., Müller, D. B., ... & Liu, G. (2022). Battery technology and recycling alone will not save the electric mobility transition from future cobalt shortages. *Nature communications*, *13*(1), 1341.

59 Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. In *Temporal Boundaries of Law and Politics* (pp. 72–87). Routledge.

Mikäli kiihtyvyyden kehällä eri osa-alueet kiihtyvät eri tahdissa, seurauksena on Rosan mukaan eritahdistuminen (*desynchronization*) ja siitä seuraavat haitat. Tarve palauttaa yhteiskunnan eri osa-alueet samaan tahtiin (*resynchronization*) ennen kuin eritahtisuudesta aiheutuvat haitat kasvavat liian suuriksi altistaa ratkaisuille, joiden vaikutuksia ei välttämättä pystytä aikapaineessa selvittämään.[60] Sekä eritahdistuminen itsessään että pyrkimys sen nopeaan korjaamiseen altistavat siis haitoille, joita voidaan arvioida yhteiskunnallisten haittojen teorian kautta. Myös esimerkiksi valtionhallinnon kasvavaa tukeutumista algoritmeihin voidaan pitää yhtenä yrityksenä vastata kiihtyvien yhteiskuntien vaatimuksiin[61].

Eritahdistumisesta voi seurata monenlaisia yhteiskunnallisia haittoja. Globaalit kriisit ilmastonmuutoksesta aseellisiin konflikteihin voidaan nähdä eritahtisen kiihtyvyyden seurauksina[62]. Monien konfliktien taustalla on kasvava kulutus, johon luonnonvarat eivät riitä: seurauksena osa maailmasta kärsii kohtuuttomasti luonnon kantokyvyn ylittymisen seurauksista ja altistuu siitä seuraaville luonnonmullistuksille ja yhteiskunnallisille kriiseille, jotka altistavat konflikteille ja kamppailulle hupenevista resursseista. Kriisien seurauksena yhä suurempi määrä ihmisiä on joutunut pakenemaan valtiosta toiseen.

Kriisien seurauksena eritahdistumisen ongelmat voivat kertaantua ajan kuluessa. Esimerkiksi vuonna 2015 Suomessa, kuten muuallakin Euroopassa, eri syistä turvapaikkaa hakevien henkilöiden määrä moninkertaistui aiemmasta. Maahanmuuttovirasto ei kyennyt tehostamaan eli kiihdyttämään toimintaansa vastaamaan uutta tilannetta, mikä suisti viraston toiminnan ongelmiin. Hallinnolliset ongelmat aiheuttivat haittoja viraston asiakaskunnalle, muun muassa turvapaikanhakijoille, kuten osatutkimuksessa 3 esitetään. Eritahdistumisen ongelmia pyrittiin hallitsemaan paitsi kiihdyttämällä Maahanmuuttoviraston toimintaa muun muassa automaation avulla, myös jarruttamalla Suomeen saapumisen mahdollisuuksia.

Kriittiset algoritmitutkijat argumentoivat kuitenkin laajasti, että teknologisilla ratkaisuilla ei ole mahdollista korjata yhteiskunnallisia ongelmia, jotka pohjaavat yhteiskunnan rakenteelliselle epäoikeudenmukaisuudelle. Päinvastoin, monet näkevät etenkin tekoälyteknologiat rakenteellisen epäoikeudenmukaisuuden ilmentymänä[63] ja siten epäoikeudenmukaisuutta toisintavina ja vahvistavina ilmiöinä[64]. Mo-

60 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

61 Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1–18.

62 Rosa, H. (2015). Escalation: The crisis of dynamic stabilisation and the prospect of resonance. Sociology, capitalism, critique, 322–347.

63 Mann, M. (2020). Technological politics of automated welfare surveillance: Social (and data) justice through critical qualitative inquiry. *Global Perspectives*, *1*(1).

64 McQuillan, D. (2022). Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press.

dernien, kapitalististen yhteiskutien rakenne ja yhteiskunnalliset järjestelmät määrittävät pitkälti, missä ja minkälaista teknologiaa voidaan hyödyntää. Väitettä on helppo ymmärtää. Kun vallassa olevilla on parhaat mahdollisuudet vaikuttaa teknologiseen kehitykseen, heidän intressissään on kehittää teknologioita suuntaan, joka ei ainakaan heikennä heidän asemaansa. Tällöin riskinä on, että haavoittuvassa asemassa olevat joutuvat entistä heikompaan asemaan. Riski on todellinen ja se tulee ottaa vakavasti. Argumentaatiossa, jossa algoritmiset teknologiat nähdään väistämättä ja ehdoitta rakenteellista epäoikeudenmukaisuutta syventävinä järjestelminä, ei kuitenkaan huomioida teknologisen kehityksen ja sen sääntelypyrkimysten taustalla käytävää jatkuvaa kamppailua siitä, mihin pyritään, mitä ja keitä otetaan huomioon, tai mitä arvoja vahvistetaan ja tuotetaan[65].

Vaikka algoritmisten teknologioiden ei arvioitaisi deterministisesti toisintavan ja vahvistavan yhteiskunnallista epäoikeudenmukaisuutta, kuten esimerkiksi Mann[66] argumentoi, on myös selvää, että jatkuva tehostaminen, lisääminen ja nopeuttaminen ei rajallisessa maailmassa ole loputtomiin mahdollista. Kun algoritmiset teknologiat lisäävät kiihdyttämisen mahdollisuuksia entisestään, niiden jatkuvasti lisääntyvä käyttöönotto vie yhteiskuntia lähemmäksi pistettä, jossa jotkin yhteiskunnan osa-alueet eivät enää pysty nopeuttamaan tai tehostamaan toimintaansa, mikä johtaa yhä pahenevaan eritahdistumiseen. Rosa argumentoikin, että moderneilla yhteiskunnilla on pääpiirteisesti kaksi vaihtoehtoa: joko pysyä jatkuvasti kiihtyvän kehityksen mukana tavalla tai toisella, mikä mitä todennäköisimmin johtaisi maapallon kantokyvyn ylittymiseen ja ekokatastrofiin, tai hylätä kiihtyvyyden logiikka, minkä seurauksena valtiot todennäköisesti suistuisivat köyhyyteen, yhteiskuntajärjestelmät joutuisivat kriiseihin, ja pahimmillaan seuraisi maailmanlaajuinen sota[67].

Yhteiskuntien perustavanlaatuinen muutos vaikuttaa ainoalta keinolta irrottautua modernin yhteiskunnan välttämättömästä tarpeesta kiihdyttää kohti omaa tuhoaan. Yhteiskunnan perustavanlaatuinen muutos nähdään myös yhteiskunnallisten haittojen tutkimuksen parissa välttämättömänä[68]: kun haitat syntyvät erilaisten yhteiskunnallisten prosessien seurauksena, näihin prosesseihin vaikuttaminen on tehokkain tai jopa ainut tapa vähentää haittojen ilmenemistä. Vaikka perustavanlaatuinen yhteiskuntajärjestelmän muutos olisi lopulta väistämätön, haittojen muodostumisen mekaniikan ymmärtäminen ja transformaation ajallisen ulottuvuuden hahmottaminen voi-

65 McCarthy, D. R. (2013). Technology and 'the international'or: How I learned to stop worrying and love determinism. *Millennium*, *41*(3), 470–490.

66 Mann, M. (2020). Technological politics of automated welfare surveillance: Social (and data) justice through critical qualitative inquiry. *Global Perspectives*, *1*(1).

67 Rosa, H. (2015). Escalation: The crisis of dynamic stabilisation and the prospect of resonance. Sociology, capitalism, critique, 322–347.

68 Canning, V., & Tombs, S. (2021). From social harm to zemiology: A critical introduction. Routledge. 49.

vat auttaa kehittämään vaihtoehtoisia toimintamuotoja, joiden avulla kiihtymisen negatiivisia lieveilmiöitä voitaisiin hallita ja siten lisätä oikeudenmukaisuutta ja inhimillistä kukoistusta. Tiedon lisääminen ja vaihtoehtoisten toimintatapojen esiin tuominen ovat myös välttämättömiä askeleita ennen kuin perustavanlaatuinen muutos voi olla mahdollinen.

1.3 Tutkimuskysymykset, menetelmä ja työn rakenne

Tässä väitöskirjassa algoritmisia teknologioita ja niiden haittapotentiaalia tarkastellaan useiden menetelmien avulla. Osatutkimuksissa on hyödynnetty niin empiirisiä kuin teoreettisia lähestymistapoja, ja yhteenvedossa osatutkimusten johtopäätelmien pohjalta käsitellään algoritmisten haittojen ominaispiirteitä ja haittojen syntymiseen vaikuttavia seikkoja sekä arvioidaan tekoälylainsäädännön vaikuttavuutta tekoälyyn kiinnittyvien haittojen hallinnan työkaluna.

Osatutkimuksista ensimmäisessä hyödynnetään tapausanalyysejä, joiden avulla arvioidaan, miten tekoälyjärjestelmät muuttavat yhteiskunnallisten haittojen typologiaa. Toisessa osatutkimuksessa tukeudutaan kriittisen diskurssianalyysiin (KDA), ja analyysin kohteena ovat valittu joukko lausuntoja, joita annettiin julkishallinnon automaattiseen päätöksentekoon liittyvän lainsäädäntöprosessin aikana. KDA:n historiallinen suuntaus (*discourse-historical approach*)[69] auttaa arvioimaan esiin nousseita diskursseja suhteessa yhteiskunnan kehitykseen – ja erityisesti Rosan[70] esiin nostamaan yhteiskunnalliseen kiihtyvyyteen. Kolmas osatutkimus taas nojaa sisällönanalyysiin, jonka avulla arvioidaan tietyissä tekoäly- ja maahanmuuttopoliittisissa dokumenteissa esiintyviä tekoälytavoitteita ja ymmärrystä mahdollisista haitoista ja riskeistä. Asetelma antaa mahdollisuuden nostaa esiin eri poliittisten sektorien ristiriitaisuuden, tekoälyn ja haittojen vaillinaisen ymmärryksen sekä valtionpolitiikan siiloutumisen mukanaan tuomien ongelmien laajuuden. Osatutkimusten tarkemmat tiivistelmät löytyvät luvusta kaksi.

Väitöskirjani yhteenvedon tavoitteena on osatutkimuksissa tehtyjen havaintojen ja johtopäätösten innoittamana lisätä ymmärrystä siitä, minkälaisia haittoja algoritmisiin teknologioihin liittyy, mihin järjestelmien ominaisuuksiin riskit ja haittapotentiaali kiinnittyvät, sekä miten EU:n tekoälyasetus ja kansallisen tason sääntely mahdollistavat potentiaalisten haittojen hallinnan. Väitöskirjan keskiössä ovat seuraavat kysymykset:

69 Wodak, R. (2015). Critical Discourse Analysis, Discourse-Historical Approach. In *The International Encyclopedia of Language and Social Interaction*, 1–14.

70 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

1) Kun algoritmisten teknologioiden vaikutuksia tarkastellaan yhteiskunnallisten haittojen teorian valossa, minkälaisia i) yksilöihin, ii) ryhmiin ja iii) yhteiskuntaan kohdistuvia haittoja on tunnistettu?
2) Miten algoritmisten teknologioiden ominaisuudet vaikuttavat erilaisten haittojen syntyyn?
3) Miten oikeudellinen sääntely pyritään tällä hetkellä järjestämään kansallisesti ja EU-tasolla, ja miten sääntelyratkaisut mahdollistavat algoritmisten teknologioiden potentiaalisesti aiheuttamien yhteiskunnallisten haittojen hallinnan?

Väitöskirjani aihepiiri on laaja ja siihen kiinnittyy kysymyksiä monilta tieteenaloilta. Tämä on johdanto-osuudessa selkeästi nähtävillä. Kysymyksiin vastaaminen vaatii monien eri teorioiden soveltamista ja myös monitieteistä tutkimusotetta. Vaikka taustani on oikeustieteessä, olen pyrkinyt laventamaan tutkimusta yli perinteisen oikeustieteen rajojen. Tutkimuksessani lähestynkin niin sosiologian, filosofian kuin politiikantutkimuksen kysymyksenasetteluja. Tämä vaatii huomattavaa tasapainoilua: yhtäältä olen joutunut rajaamaan syvyyttä, jolla tuon esiin eri näkökulmia, ja toisaalta olen joutunut sovittamaan yhteen tutkimusperinteitä, joita ei useinkaan käsitellä yhdessä. Vaikka pyrkimykseni on tuoda eri tieteenaloja lähemmäs toisiaan, on kuitenkin huomattava myös, että taustani ollessa oikeustieteessä ja oikeussosiologiassa, tarkastelen näistä aloista etääntyviä teorioita ja näkökulmia mahdollisesti totutusta poikkeavalla tavalla.

Ymmärrän algoritmisen transformaation muutosvoimana, joka kiinnittyy osaksi kiihtyvää, modernia yhteiskuntaa. Muutoksen analyysissä hyödynnän paitsi yhteiskunnallisen kiihtyvyyden teoriaa myös rakenteistumisteoriaa, kun taas haittoja analysoin kriminologisesta perinteestä irtautuneen yhteiskunnallisten haittojen teorian pohjalta. Näiden lisäksi tukeudun kriittisen tekoälytutkimuksen tutkimusperinteeseen, kun arvioin tekoälyn ja algoritmisten teknologioiden vaikutuksia yksilöihin ja yhteiskuntaan. Pyrin nostamaan argumentaationi tueksi myös yhteiskuntatieteellistä näkökulmaa teknisemmän tavan arvioida algoritmisia teknologioita alan kirjallisuuteen tukeutuen. Algoritmisten teknologioiden ja niihin kytkeytyvien haittojen analyysi muodostaa pohjan oikeudelliselle arvioinnille. Toisin sanoen hyödynnän oikeustieteen ulkopuolelta kerryttämääni ymmärrystä modernista todellisuudesta, jossa tekoälyteknologiat ovat merkittävässä roolissa, kun analysoin tekoälysääntelyn mahdollisuuksia hallita algoritmisesta transformaatiosta mahdollisesti seuraavia haittoja.

Yhteiskuntatieteellinen ja oikeustieteellinen käsitys tiedosta ja sen luonteesta luonnollisesti eroavat toisistaan, kuten myös näkökulmat, joiden kautta oikeudellisia ilmiöitä tarkastellaan. Normatiivinen oikeustiede pyrkii systematisoimaan ja tulkitsemaan lainsäädäntöä, jolloin näkökulma on *oikeudensisäinen* ja tiedonintressi kiin-

nittynyt oikeuslähdeopin mukaisesti säädännäiseen oikeuteen, lainsäädännön valmistelumateriaaleihin, ratkaisukäytäntöihin ja oikeudelliseen tutkimukseen. Yhteiskuntatieteellinen tutkimus taas lähestyy oikeutta yhteiskunnallisena ilmiönä ja arvioi sen vuorovaikutusta muun yhteiskunnan kanssa *ulkoisesta* näkökulmasta. Ulkoinen näkökulma on kriittisen kriminologian ja yhteiskunnallisten haittojen tutkimuksen parissa tyypillistä. Arvioinnin kohteena ovat yhteiskunnalliset ilmiöt, joita oikeudellinen järjestelmä mahdollistaa, ylläpitää tai ruokkii.

Vaikka tässä tutkimuksessa oikeudellisella analyysilla on merkityksellinen rooli, kontribuutio painottuu kenties sitäkin vahvemmin kriittisen tekoälytutkimuksen ja erityisesti yhteiskunnallisten haittojen tutkimuksen kentälle. Tekoälyn tutkimus on nopeasti kasvava tutkimusala. Viimeisiä editointeja tehdessäni loppukeväästä 2024 Google Scholarin arvosteluartikkelihaku hakusanoilla ”artificial intelligence” ja ”regulation” tuotti yli 5000 osumaa vuoden 2024 ajalta. ”Artificial intelligence” ja ”justice” toi 1600 hakutulosta. Onkin selvää, että valtava määrä tutkimusta on väistämättä rajautunut tämän työn ulkopuolelle. Työni pääasiallisena tutkimuskohteena ja painopisteenä on tekoälyjärjestelmien aiheuttamat potentiaaliset haitat ja niiden hallitseminen modernin yhteiskunnan puitteissa, ja lainsäädäntö hahmottuu työkaluna, joka parhaimmillaan rajoittaa ja ennaltaehkäisee, huonoimmillaan mahdollistaa ja kiihdyttää haittojen ilmenemistä. Arvioinnin ja analyysin kohteena ei siis ole tekoälyyn liittyvä lainsäädäntö suljettuna järjestelmänä; puhtaasti lainopillisen tarkastelun sijasta pyrin arvioimaan, miten tekoälyn sääntely vaikuttaisi onnistuvan modernissa yhteiskunnassa ilmenevien algoritmisten haittojen rajoittamisessa ja hallitsemissa.

Olen yllä kuvannut tutkimuksen lähtökohtia ja sen teoriapohjaa. Väitöskirja etenee siten, että osatutkimusten tiivistelmien (luku 2) jälkeen käsittelen yhteiskunnallisten haittojen arvioimisen tapoja ja haittoja, joita algoritmisiin teknologioihin on tunnistettu liittyvän (luku 3). Sen jälkeen arvioin, miten algoritmisten teknologioiden ominaisuudet selittävät haittojen syntymistä (luku 4). Näiden jälkeen siirryn käsittelemään algoritmisiin teknologioihin keskittyvää lainsäädäntöä EU- ja kansallisella tasolla (luku 5). Tutkimusasetelman taustalla vaikuttaa kysymys siitä, millaiseen yhteiskuntaan olemme tekoälyn aikakaudella astumassa, ja onko se sitä, mitä yhteiskunnaltamme haluamme – ja mikäli emme, mitä mahdollisuuksia meillä olisi lainsäädännön keinoin muuttaa suuntaa. Väitöstutkimukseni keskeiset tulokset ja niihin perustuvat johtopäätelmät esitän luvussa 6.

2 Osatutkimusten tiivistelmät

Tässä luvussa käyn tiiviisti läpi osatutkimusten tutkimusasetelmat, menetelmät, keskeisen sisällön ja tulokset. Osatutkimukset muodostavat jatkumon niin ajankuvauksen kuin käsiteltyjen teemojen osalta. Vaikka niiden julkaisuajat ovat lähellä toisiaan, ne on kirjoitettu useamman vuoden aikana – ensimmäisen artikkelin osalta työ alkoi vuonna 2018, ja viimeinen artikkeli valmistui vuonna 2022. Tänä lyhyenäkin aikana tekoälyn ja automaation käytön tavat ja mahdollisuudet yhteiskunnassa muuttuivat valtavasti. Viimeisen artikkelin julkaisemisen jälkeen niin sanottu generatiivinen tekoäly, jota käsittelen tekoälyn luokittelun yhteydessä enemmän, on yhä kiihdyttänyt tekoälykeskustelua ja algoritmisen transformaation vauhtia. Tämän luvun tarkoituksena on paitsi tiivistää osatutkimusten anti myös osoittaa osatutkimusten kuvauksen avulla, kuinka algoritmisen transformaation kiihtyminen on näkynyt yhteiskunnassa, ja siten perustella, minkä takia tarve tutkimukselle on edelleen valtava.

Osatutkimuksissa on hyödynnetty eri menetelmiä ja erilaisia aineistoja. Ensimmäinen artikkeli sijoittuu globaaliin kontekstiin. Siinä keskitytään yhteiskunnallisten haittojen typologiaan ja argumentoidaan, että algoritmiset teknologiat vaikuttavat haittojen muodostumiseen. Toisessa ja kolmannessa osatutkimuksessa tarkastellaan Suomea ja algoritmista päätöksentekoa julkisella sektorilla. Artikkeleissa rakennetaan laadullisen tutkimuksen keinoin kuvaa siitä, miten tekoäly ja automaatio ymmärretään Suomen valtionpolitiikan ja lainsäädännön kontekstissa, ja minkälaisissa tilanteissa niiden hyödyntämiseen pyritään.

Kuvaan ensimmäisessä alaluvussa osatutkimusten tutkimusasetelmat ja menetelmät. Sitä seuraavissa alaluvuissa tiivistän osatutkimus kerrallaan tutkimusten keskeisen sisällön ja tutkimustulokset.

2.1 Tutkimusasetelmat

Ensimmäisessä osatutkimuksessa[71] tarkastelin kollegoideni kanssa kolmea laajasti tunnettua tapausta, joissa algoritmiset teknologiat ovat olleet osallisena haittojen aiheuttamisessa: Yhdysvaltojen MiDAS-petoksentunnistusohjelmaa, vuoden 2010 osakemarkkinoiden ”flash crashiä” ja Cambridge Analytican vaalivaikuttamisskandaalia Yhdysvaltojen presidentinvaaleissa vuonna 2016. Tapauksista oli kirjoitettu paljon jo ennen tutkimustamme. Aiemmat tutkimukset toimivat analyysin pohjana ja antoivat mahdollisuuden arvioida nimenomaan algoritmisten teknologioiden vaikutuksia yhteiskunnallisten haittojen muodostumisessa. Tarkastelemalla laajasti tutkittuja ja hyvin tunnettuja tapauksia ja peilaamalla niitä yhteiskunnallisten haittojen ja kriittisen algoritmitutkimuksen kirjallisuuteen, pystyimme arvioimaan algoritmisten teknologioiden haittapotentiaalia ja niiden ominaispiirteitä.

Analyysissa olimme ensi sijassa kiinnostuneita algoritmien aikaansaamista *muutoksista* tunnettujen haittojen muodostumisessa ja vaikutuksissa. Näin ollen keskityimme vertaamaan valittujen tapausten aikaansaamien haittojen ominaispiirteitä perinteisesti yhteiskunnallisten haittojen tutkimuksessa tunnistettujen haittojen ominaispiirteisiin. Artikkelissa pystyimme osoittamaan, että algoritmisaatiolla ja tekoälyn lisääntyvällä hyödyntämisellä on merkitystä yhteiskunnallisten haittojen ilmenemisessä, mitä ennen artikkelin ilmestymistä ei juurikaan ollut tutkittu.

Osatutkimuksissa 2[72] ja 3[73], jotka kirjoitin yhdessä Hanna Malikin kanssa, hyödynsimme laadullisia menetelmiä. Toisen osatutkimuksen menetelmäksi valitsimme diskurssianalyysin, joka pohjautuu ajatukselle siitä, että kieli paitsi kuvaa todellisuutta, myös rakentaa ja muokkaa sitä vahvistaen, heikentäen, luoden ja tuhoten erilaisia tapoja nähdä ja tulkita ympäröivää maailmaa[74]. Tarkastelimme kriittisen diskurssianalyysin keinoin tekoälydiskursseja Suomen hallinnollisen päätöksenteon automaation mahdollistavan yleislain valmistelun yhteydessä. Sosiaalisen konstruktionismin hengessä uskomme, että tapa, jolla tekoälystä ja automaatiosta puhutaan, vaikuttaa siihen, miten tekoäly nähdään – ja se, miten tekoäly nähdään, vaikuttaa siihen, miten sitä koetaan tarpeelliseksi säännellä.

71 Malik, H. M., Viljanen, M., Lepinkainen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

72 Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

73 Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 270–291.

74 Wodak, R., & Meyer, M. (eds). (2001). *Methods of Critical Discourse Analysis*. London: SAGE Publications, Ltd. https://doi.org/10.4135/978085702 8020.
Fairclough, N. (1996). *Discourse and Social Change*. Cambridge: Polity Press.

Osatutkimuksessa selvitimme, miten tekoälyä ja automaatiota kuvataan valitussa aineistossa: mitä riskejä niihin liitetään, mitä hyötyjä niistä hahmotetaan seuraavan, minkälaista tulevaisuutta näkemyksissä rakennetaan, ja miten esitetyt näkökulmat sopivat yhteen hyvää hallintoa koskevien perustuslaillisten velvoitteiden kanssa. Rajasimme tutkittavan aineiston jokseenkin kapeaksi. Rajaus perustuu pyrkimykseen keskittyä nimenomaan sellaisiin lausuntoihin, jotka tutkimusten perusteella vaikuttavat vahvimmin lainsäädännön muodostumiseen. Muiden muassa Helminen ja Alvesalo-Kuusi[75] ovat argumentoineet lainsäädännön merkittävimpien linjausten hahmottuvan nimenomaan lainsäädännön valmisteluvaiheessa, mihin pääsevät vaikuttamaan vahvimmin juuri lainvalmistelutyöryhmän jäsenet. Analysoitavaksi aineistoksi valitsimmekin lausunnot, jotka lainvalmistelutyöryhmään valitut tahot antoivat lakihankkeen ensimmäisellä lausuntokierroksella heti oikeusministeriön esivalmistelun jälkeen.

Hyödynsimme historiallis-diskursiivista lähestymistapaa[76], mikä mahdollisti analyysin yhdistämisen historialliseen ja yhteiskunnalliseen kontekstiinsa. Diskurssianalyysin tulosten jäsentely Rosan yhteiskunnallisen kiihtyvyyden teorian[77] valossa antoi mahdollisuuden arvioida diskurssien taustalla olevia tekijöitä sekä niiden mahdollisia vaikutuksia.

Kolmannessa artikkelissa pyrimme rakentamaan holistista kuvaa siitä, miten yhtäältä tekoälypolitiikan, toisaalta maahanmuuttopolitiikan kontekstissa käsiteltiin tutkimuksen tekemisen aikana automaation mahdollisuuksia ja riskejä, ja miten erilaisia haittoja ymmärrettiin. Jotta lukijalla olisi mahdollisuus asettaa analyysi kontekstiinsa, selvitimme artikkelissa ensin Maahanmuuttoviraston päätöksenteon kriisiytymistä, kriisin ratkaisupyrkimyksiä ja automaatiosuunnitelmien kehittymistä vuoden 2015 niin kutsutun pakolaiskriisin jälkeen. Tilannekuvan rakentamisessa hyödynsimme laajasti erilaisia lähteitä: paitsi tieteellisiä ja journalistisia artikkeleita myös lehdistötiedotteita sekä muita julkisesti saatavilla olevia dokumentteja.

Analyysin kohteeksi valitsimme merkittävimmät poliittiset ja lainsäädännölliset dokumentit maahanmuutto- ja tekoälypolitiikan alueilta[78]. Laajan aineiston sisällönanalyysin perusteella halusimme luoda holistisen kuvan siitä, minkälaisia painotuksia eri politiikan alueilla nousi esille, minkälaisiin seikkoihin kiinnitettiin huo-

75 Helminen, M., & Alvesalo-Kuusi, A. (2017). Advocating the 'Good' Criminal Justice System. The Involvement and Ideas of Civil Society Organisations in Formulating Finnish Criminal Policy. *Retfærd. Nordic Journal of Law and Justice* 40 (2), 3–24.

76 Wodak, R. (2015). Critical Discourse Analysis, Discourse-Historical Approach. In *The International Encyclopedia of Language and Social Interaction*, 1–14.

77 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

78 Katso tarkka aineisto: Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3). 272.

miota ja miten eri sektoreiden linjaukset sopivat yhteen tai olivat keskenään ristiriidassa. Etenkin, kun huomioidaan, kuinka algoritminen transformaatio vaikuttaa koko yhteiskuntaan ja yli poliittisten sektoreiden, pidimme tärkeänä tutkia automatisaatio- ja tekoälytavoitteita suhteessa eri käyttökonteksteihin. Vertailemalla tekoälypolitiikassa ja maahanmuuttopolitiikassa esiin nostettuja haittoja, riskejä ja mahdollisuuksia pystyimmekin osoittamaan poliittisten tavoitteiden ristiriitoja ja päällekkäisyyksiä sekä tunnistamaan teemoja, joita politiikassa ei riittävästi käsitellä. Kiinnitimme analyysimme Pembertonin vähiten haitallisten yhteiskuntamuotojen (harm reduction regimes[79]) teoreettiseen viitekehykseen.

Taulukko 1. Osatutkimusten tutkimuskysymykset ja menetelmät.

	Tutkimuskysymykset	**Data/lähdeaineisto**	**Lähestymistapa**
Osatutkimus 1	Miten algoritmiset teknologiat vaikuttavat yhteiskunnallisten haittojen syntyyn, kun niitä tarkastellaan monimutkaisessa, modernissa sosioteknisessä ympäristössä?		Yhteiskunnallisten haittojen teoria
Osatutkimus 2	Minkälaisia tekoälydiskursseja aineistosta voidaan havaita, ja minkälaisia sääntelyvaihtoehtoja niiden voidaan katsoa perustelevan? Miten nämä diskurssit vaikuttavat Suomen perustuslain tulkintaan?	Lainvalmistelutyöryhmän antamat lausunnot hallinnon automaattisen päätöksenteon yleislainsäädäntelytarvetta arvioivasta arviomuistiosta[80]	Kriittinen diskurssianalyysi, historiallis-diskursiivinen lähestymistapa
Osatutkimus 3	Miten Suomen tapa suhtautua automaattiseen päätöksentekoon sopii yhteen niiden perinteiden kanssa, jotka ovat nostaneet Suomen harm reduction regimes[81] -jaottelussa yhdeksi vähiten haitallisista valtioista? Miten erilaisten haittojen yhteydet huomioidaan poliittisissa ja lainsäädännöllisissä dokumenteissa?	Valitut poliittiset dokumentit ja lainvalmistelumateriaalit tekoäly- ja maahanmuuttopolitiikan kentältä[82]	Sisällönanalyysi

79 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

80 Ks. tarkka kuvaus aineistosta: Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707. 4–5.

81 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

82 Ks. tarkka kuvaus aineistosta: Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3). 272.

Seuraavissa alaluvuissa käsittelen osatutkimuksia ja niiden tuloksia yksityiskohtaisemmin.

2.2 Osatutkimus 1: Dynamics of Social Harms in an Algorithmic Context

Kuten yllä tiiviisti kuvattiin, ensimmäisessä osatutkimuksessa, *Dynamics of Social Harms in an Algorithmic Context,* selvitimme, millä tavoin algoritmiset teknologiat olivat vaikuttaneet ja yhä vaikuttavat systeemitasoisten, laajalle levittäytyvien haittojen muodostumiseen nykypäivän monimutkaisissa yhteiskunnissa, joissa älysovellukset ovat osa arkea. Keskityimme yhteiskunnallisten haittojen teorian tutkimuskirjallisuudessa laajasti tutkittuihin haittoihin: ekonomisiin, taloudellisiin, psykologisiin[83] sekä tunne- ja kulttuurisiin[84] haittoihin, turvallisuuden heikentämiseen, eri ryhmien väärin tunnistamiseen (*misrecognition*)[85], syrjäytymiseen eli ekskluusioon, sekä autonomian heikentämiseen[86].

Valitsimme artikkelin tutkimuskohteeksi kolme laajasti tunnettua tapausta, joissa algoritmiset teknologiat todistetusti olivat olleet osallisina haittojen aiheuttamisessa: Yhdysvaltojen MiDAS-petoksentunnistusohjelman, vuoden 2010 osakemarkkinoiden ”flash crashin” ja Cambridge Analytican vaalivaikuttamisskandaalin Yhdysvaltojen presidentinvaaleissa vuonna 2016. Valitut tapaukset eivät tietenkään edusta täydellisesti algoritmisaation mukanaan tuomia muutoksia. Tämän takia emme pyrkineet analyysissa tunnistamaan kaikkia mahdollisia haittoja, joita algoritmit voivat tuoda mukanaan, vaan keskityimme vertaamaan valittujen tapausten aikaansaamien haittojen ominaispiirteitä yhteiskunnallisten haittojen tutkimuksessa tunnistettujen, niin sanotusti analogisten haittojen ominaispiirteisiin. Tapausten analyysin pohjalta esitämme, että algoritmit vaikuttavat merkittävästi yhteiskunnallisten haittojen muodostumiseen: tekoälyteknologioiden vuoksi haitat voivat systematisoitua, ja samalla teknologiat voivat voimistaa haittoja, nopeuttaa niiden leviämistä ja vaikeuttaa niiden syiden ja syntymekanismien tunnistamista.

Ensimmäiseksi analysoimme Michiganissa vuonna 2013 käyttöön otettua algoritmia, jolla pyrittiin tunnistamaan avustuspetoksia. Algoritmia kutsuttiin nimellä MiDAS: Michigan Integrated Data Automated System. Kun algoritmi oli ollut käy-

83 Tombs, S. (2019). Grenfell: the unfolding dimensions of social harm. *Justice, Power and Resistance 3/1*. 61–88.

84 Alvesalo, A. (1999). Meeting the expectations of the local community on safety—what about white-collar crime? Konferenssiesitys *27th Annual Conference of the European Group for the Study of Deviance and Social Control*. Liettua, 2.–5.9.1999.

85 Yar, M. (2012). Critical criminology, critical theory and social harm. In S. Hall, S. Winlow, (eds). *New Directions in Critical Theory*. 52–63. Routledge.

86 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

tössä kaksi vuotta, havaittiin, että se epäsi avustukset 13 prosenttiyksikköä useammin kuin avustusten myöntämisestä aiemmin vastanneet ihmiset. Lisäksi se merkitsi kaksinkertaisen määrän hakemuksia mahdollisesti petollisiksi. Kyse ei ollut tarkkuuden tai tehokkuuden lisääntymisestä. Sen sijaan kyseessä oli puhtaasti algoritmin virheellinen toiminta: tutkimuksissa on arvioitu, että petosepäilyjen kohdalla virheellisten päätösten määrä olisi ollut noin 85 prosenttia[87].

Arviomme mukaan haitat, jotka syntyivät MiDAS-algoritmin hyödyntämisestä, olivat verrattain tyypillisiä yhteiskunnallisia haittoja. Algoritmi esti oikeutettujen avustusten myöntämisen, mistä seurasi taloudellisia haittoja. Petosepäily käynnisti rikosprosessin. Taloudelliset haitat ja rikosprosessin aikaansaama stigma rajoittivat ihmisten autonomiaa ja mahdollisuuksia osallistua sosiaaliseen kanssakäymiseen, lisäsivät epävarmuutta ja sysäsivät virheellisten päätösten kohteet yleisen epäluulon alaisiksi. Stigma saattoi johtaa myös syrjäytymiseen tai ekskluusioon[88].

Samanlaisia haittoja syntyy myös ihmisten tekemien virheellisten päätösten seurauksena. Algoritmiavusteisella päätöksenteolla on kuitenkin ominaispiirteitä, jotka muuttavat haittojen syntymekaniikkaa. Kun päätöksenteko on automatisoitu, jokaiseen tapaukseen sovelletaan aina samoja päätösperusteita. Mikäli päätösperusteet ovat virheellisiä, se näkyy systemaattisesti jokaisessa tehdyssä päätöksessä. Automaattisella päätöksenteolla on siis suuri potentiaali automatisoida ja systematisoida virheitä. Kun vertasimme algoritmista päätöksentekoa ihmisvetoiseen päätöksentekoon, huomasimme, että ihmisvetoisessa päätöksenteossa yksi suojaava seikka on se, että päätöksentekovastuu on silloin hajautunutta ja päätöksentekojärjestelmässä on sen myötä varmuuskerroksia: ihmiset tekevät päätöksiä eri perusteilla ja pystyvät myös havaitsemaan poikkeavat päätöksentekotavat. Automatisoidun päätöksenteon tapauksessa näin ei MiDAS-tapauksen perusteella ole. Virheellisiin päätösperusteisiin puuttuminen viivästyi mahdollisesti ainakin osaltaan sen takia, että hajautuksen mukanaan tuomat varmuuskerrokset puuttuivat. Ongelmat syvenivät, sillä kun virheiden määrä lisääntyi, muutoksenhakuinstanssit kuormittuivat suhteettomasti, mikä hankaloitti oikeuden saamista ja pidensi virheiden aikaansaamien haittojen kestoa.

Seuraavaksi käsittelimme osakemarkkinoita. Tapaus sai alkunsa, kun toukokuussa 2010 Yhdysvaltojen osakemarkkinoilla pörssikurssit romahtivat yllättävästi. US Commodity Futures Trading Commissionin ja US Securities & Exchange Com-

87 Shaefer, H.L., & Grey, S. (2015). Letter to U.S. Department of Labor – Michigan Unemployment Insurance Agency: Unjust Fraud and Multiple-Determinations.

88 Ekskluusio juontuu englannin kielen termistä exclude, joka usein käännetään suomeksi syrjäytymiseksi. Exclude viittaa kuitenkin ennemmin *syrjäyttämiseen*: aktiiviseen toimintaan, joka sulkee jonkun ulos jostakin. Tämän vuoksi itse suosin anglismia. Siinä missä syrjäytyminen antaa kuvan muiden toiminnasta riippumattomasta seikasta, ekskluusioon sisältyy ajatus siitä, että syrjäytyminen usein aiheutetaan.

missionin raportti[89] osoitti, että romahdus johtui nykyisten osakemarkkinoiden ominaisuuksista: yksittäinen, riittävän suuri ostotapahtuma voi suistaa osakekurssit yhtäkkisesti romahtamaan.

Ensi silmäyksellä osakemarkkinat ja algoritmiset haitat eivät välttämättä näytä liittyvän yhteen. Nykypäivää osakkeiden pörssikauppa on kuitenkin pitkälti automatisoitua. Kaupankäynnissä hyödynnetään algoritmeja, joiden toiminta voi vaikuttaa haittojen syntyyn. Tämä voi johtua eri tekijöistä. Yhtäältä algoritmien koodaamisessa voi tapahtua virheitä, ja toisaalta myös virheettömästi koodattu algoritmi voi rakentua väärien tai puutteellisten arvioien varaan ja siis toimia näiden arvioiden takia epätoivotulla tavalla. Algoritmit toimivat äärimmäisen nopeasti, ja niiden toimintaan voi olla lyhyen aikajänteen takia käytännössä mahdotonta puuttua.

Tutkimukset osoittavat, että vuoden 2010 pörssiromahduksessa algoritmeilla oli suuri merkitys: erään yrityksen algoritmi oli koodattu hinnasta välittämättä ostamaan joka minuutti 9 % saatavilla olevista S&P E-mini futuurisopimuksista. Futuurisopimuksilla on rajattu markkinalikviditeetti[90] ja ostajia ja myyjiä on vähän, minkä seurauksena algoritmi päätyi maksamaan sopimuksista koko ajan korkeampaa hintaa[91]. Hintojen äkillinen nousu voi merkitä sitä, että osakemarkkinat ovat ongelmissa. Markkina-algoritmit vastasivatkin ongelmia indikoivaan hintojen nousuun tavalla, joka sysäsi romahdukseen johtaneen ketjureaktion liikkeelle. Algoritmien valtava toimintanopeus esti ihmisten mahdollisuudet arvioida tilannetta ja puuttua siihen[92].

Romahdus jäi onneksi lyhytaikaiseksi. Tapauksen avulla pystyimme kuitenkin osoittamaan algoritmisen pörssikaupan potentiaaliset vaarat: romahdus olisi voinut syöstä maailmantalouden globaaliin taantumaan. Algoritmien nopeus ja jatkuvasti kiihtyvät markkinatalouden prosessit lyhentävät ongelmien ilmenemisen aikajännettä ja kaventavat aikaikkunaa, jolloin ongelmiin voidaan reagoida. Kun tähän lisätään se, että algoritmit reagoivat automaattisesti markkinoiden toimintaan ja saattavat laukaista vaikeasti ennakoitavia vuorovaikutusketjuja, potentiaaliset ongelmat eivät pysähdy valtioiden rajoihin. Algoritmisen osakekaupan nopeus ja algoritmien automaattinen reagoiminen osakekaupassa tapahtuviin muutoksiin siis lisäävät hait-

89 U.S. Commodity Futures Trading Commission, & U.S. Securities & Exchange Commission. (2010). Finding regarding the market events of May 6. 2010. Report of the staffs of the CTFC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. www.cftc.gov

90 Käytännössä siis markkinahinnat muuttuvat herkästi kaupankäynnin seurauksena. Ks. esimerkiksi Manninen, O. (2016). Markkinalikviditeetistä, sen merkityksestä ja kestävyydestä. Suomen pankin ajankohtaisia artikkeleita taloudesta.

91 U.S. Commodity Futures Trading Commission, & U.S. Securities & Exchange Commission. (2010). Finding regarding the market events of May 6. 2010. Report of the staffs of the CTFC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues: 2. www.cftc.gov

92 Akansu, A. N. (2017). The flash crash: a review. *Journal of Capital Markets Studies*.

tojen syntymisen mahdollisuuksia, ja potentiaaliset haitat voivat levitä valtavalla nopeudella huomattavan laajalle alueelle.

Kolmanneksi tutkimuskohteeksi valitsimme kohdennetun vaikuttamisen sosiaalisen median alustoilla. Sosiaalinen media mahdollistaa muun muassa suosittelualgoritmien avulla valikoidun sisällön esittämisen soveltuvaksi arvioidulle joukolle, mikä tekee siitä merkittävän mielipidevaikuttamisen ympäristön. Yksi tunnetuimpia toimijoita aiheen parissa on ollut Cambridge Analytican (CA), poliittisen konsultoinnin yritys, joka vaikutti esimerkiksi Trumpin presidentinvaalikampanja taustalla Yhdysvalloissa vuonna 2016, Britannian Leave EU -kampanjassa ja monissa pienemmissä kampanjoissa ympäri maailman. Lähdimmekin selvittämään, minkälaisia haittoja CA:n vaalivaikuttamisen kampanjoista seurasi, ja mikä osuus algoritmeilla oli niiden syntymisessä.

Kun CA:n toimintaa on tutkittu, on pystytty selvittämään, että CA pyrki profiloimaan Facebook-käyttäjien persoonallisuudenpiirteitä, joiden perusteella se syötti tarkasti määriteltyä sisältöä tarkasti määritellylle käyttäjäkunnalle. Tämä oli erityisen tehokasta siksi, että suosittelualgoritmit tehostivat yrityksen toimia. Kun yritys pystyi esittämään käyttäjille haluamaansa sisältöä, se väitti voivansa lisätä polarisaatiota ja ääriajattelua ja pystyvänsä sitä kautta vaikuttamaan äänestäjien äänestyspäätöksiin. Todellisia vaikutuksia on kuitenkin äärimmäisen vaikea tutkia. Niitä on kuvattu paitsi täysin mitättömiksi[93], myös maailmaa muuttaviksi[94]. On näytetty, että kohdennetulla mainonnalla voidaan vaikuttaa ihmisten toimintaan[95], mutta samaan aikaan äänestyskäyttäytymiseen vaikuttaminen minkäänlaisin keinoin näyttäisi olevan jokseenkin vaikeaa[96].

Tapauksen olennaisin piirre yhteiskunnallisten haittojen kannalta vaikuttaisikin olevan se, kuinka hyvin algoritmit piilottavat vaikutuksia ja hämärtävät syy-seuraussuhteita. Kohdennettu sisältö sosiaalisen median alustoilla on omiaan vaikuttamaan siihen, mitä ihmiset tietävät, ja tieto taas on välttämätöntä tietoisten valintojen tekemiseksi; kun rajoitetaan, mitä tietoa ihmiset saavat, rajoitetaan myös heidän kykyään

93 Sumpter, D. (2018). Outnumbered: From Facebook and Google to Fake News and Filter-bubbles - the algorithms that control our lives. Bloomsbury Publishing.

94 Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48/2. 204–231.

95 Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Journal on Telecommunications & High Tech Law*, *13*/23. 203–216;
Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.

96 Kalla, J. L., & Broockman, D. E. (2018). The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *American Political Science Review*, 112/1. 148–166.

autonomisesti päättää, mitä he haluavat[97]. Autonomiahaittoja syntyy. Kun ihmisten saatavilla olevaan tietoon pääsee vaikuttamaan näkymättömissä, vallankäyttäjää ja vallankäytön vaikutuksia on vaikea huomata. Kun suosittelualgoritmit häivyttävät inhimillisen vaikutuksen pois näkyvistä, vallan olemassaoloa ylipäätään on äärimmäisen vaikea tutkia tai näyttää toteen.

Tapausanalyysien avulla osoitimme artikkelissamme, että algoritmisilla teknologioilla on selkeä arvolataus: siinä missä muitakin teknologioita, myös algoritmisia sovelluksia käytetään rakentamaan, muokkaamaan ja purkamaan yhteiskunnallisia rakenteita tiedostettujen ja tiedostamattomien arvovalintojen pohjalta. Optimistisuus uusien, tehokkuutta lisäävien sovellusten edessä, etenkin yhdistettynä niukkuuspolitiikkaan ja ihmistyövoiman radikaaliin vähentämiseen, avaa ovet virheiden systematisoimiseen ja keskittämiseen. MiDAS osoitti selkeästi todeksi, että algoritmisten teknologioiden käyttäminen automaattisessa päätöksenteossa voi lisätä syrjintää ja automatisoida vinoumat[98]. Toisaalta osakemarkkinoiden nopea romahdus käänsi huomion myös siihen, kuinka avuttomia ihmiset ovat valtavalla nopeudella toimivien, toistensa toimintaan reagoivien algoritmien edessä. Haittojen välttäminen systeemitasolla hauraissa järjestelmissä voi olla kiinni silkasta tuurista, kun algoritmit toteuttavat koodinsa mukaista toimintaa valtavalla nopeudella ja järkähtämättömällä logiikalla. Lisäksi, kuten CA osoitti, algoritmit hämärtävät syy-seuraussuhteita. Niiden avulla voidaan kohdentaa käyttäjille haluttua sisältöä (*microtargeting*)[99], mikä lisää manipulointimahdollisuuksia. Sosiaalinen media levittää vaikutukset laajalle. Kun algoritmit rajaavat ihmisten tiedonsaantia ja tuuppivat (*nudge*[100], *hypernudge*[101]) heitä haluttuun suuntaan, ne vaikuttavat myös ihmisten kykyyn tehdä autonomisia päätöksiä. Tiedonsaanti ja mahdollisuudet toimia vapaasti tiedon pohjalta ovat välttämättömiä paitsi yksilön autonomialle myös demokratialle. Vaikutukset voivat pahimmillaan uhata yhteiskunnan toimintaa ja oikeudenmukaisuutta[102]. On vaikeaa tai jopa mahdotonta arvioida, miten ja kuinka laajasti algoritmiset teknologiat vaikuttavat ihmisiin. Lisäksi algoritmien koodia ja toiminnan logiikkaa suojellaan usein liikesalaisuuksiin vetoamalla, jolloin myös niiden toiminnan ulkopuolinen arvioiminen ja vaikutuksiin puuttuminen on vaikeaa.

97 Pemberton, S. (2015). *Harmful Societies: Understanding Social Harm*. Policy Press. 29.

98 Eubanks, V. (2017). *Automating Inequality*. St. Martin's Press.

99 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*/2, 205395171667967.

100 Thaler, R. H. (2018). Nudge, not sludge. *Science*, *361*/6401, 431.

101 Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136.

102 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance and Society*, *15*/5. 609– 625.

Ensimmäisen osatutkimuksen keskeisin huomio on, että algoritmit vaikuttavat haittojen syntyyn sekä laadullisesti että määrällisesti, vaikka ne eivät välttämättä juurikaan muuta sitä, *minkälaisia* haittoja yhteiskunnassa ylipäätään esiintyy. Algoritmien hyödyntäminen lisää haittojen leviämisen potentiaalia ja haittojen vakavuutta. Haitoista voi kärsiä entistä laajempi joukko, sillä algoritmisessa toiminnassa virheet systematisoituvat ja leviävät nopeammin kuin ihmisjohtoisissa prosesseissa. Samaan aikaan potentiaalisten haittojen aiheuttajaa voi algoritmisessa ympäristössä olla vaikea hahmottaa, ja haittoja voi olla vaikea tunnistaa.

Artikkelimme siis osoittaa, että algoritmiset teknologiat 1) systematisoivat haittojen tuotannon ja vaikeuttavat haittojen tunnistamista ja niihin puuttumista, koska algoritminen päätöksenteko on tyypillisesti säännönmukaista ja keskitettyä, 2) voimistavat ja lisäävät haittoja, sillä digitaaliset ympäristöt ja niiden väliset yhteydet levittävät haitat nopeammin ja laajemmalle alueelle, mikä jälleen vaikeuttaa myös ihmisten mahdollisuuksia puuttua haittojen syntyyn ajoissa, ja 3) hämärtävät käsitystä haittojen olemassaolosta ja syy-seuraussuhteista sekä tekevät haittojen jäljittämisestä ja rajoittamisesta vaikeampaa, kun ympäristö muuttuu jatkuvasti monimutkaisemmaksi ja läpinäkymättömämmäksi.

2.3 Osatutkimus 2: Discourses on AI and Regulation of Automated Decision-Making

Toisen osatutkimuksen keskiössä on Suomen julkishallinnon automaattisen päätöksenteon säädösvalmistelu. Artikkelissa *Discourses on AI and Regulation of Automated Decision-Making* analysoimme Hanna Malikin kanssa Suomen lainsäädäntöprosessissa esiintyviä diskursseja tekoälystä ja automaattisesta päätöksenteosta sekä niiden suhdetta tekoälyn sääntelyvaihtoehtoihin ja Suomen perustuslaintulkintaan. Automaattinen päätöksenteko sallittiin julkishallinnossa vuoden 2023 toukokuussa laajan selvitys- ja valmistelutyön jälkeen.

Artikkelin kirjoittamisen aikaan säädösvalmistelu oli parhaillaan käynnissä. Selvitystyö oli käynnistetty sen jälkeen, kun perustuslakivaliokunta lausunnossaan 7/2019 vp oli vaatinut, että automaattisen päätöksenteon vaikutukset ja mahdollinen tarve sen yleislaintasoiselle sääntelylle olisi arvioitava ennen kuin automaattinen päätöksenteko voitaisiin julkisella sektorilla sallia. Lausunnon seurauksena oikeusministeriö alkoi työstää arviomuistiota hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista (tästedes Arviomuistio). Arviomuistio julkaistiin 6.7.2020, ja tyypilliseen tapaan siitä järjestettiin sen jälkeen lausuntokierros. Lausunnonantajien joukosta osa valittiin jatkamaan lainsäädännön valmistelua lainsäädäntötyöryhmässä.

Artikkeliamme varten tutkimme lainsäädäntötyöryhmään valittujen tahojen lausuntoja[103]. Osatutkimuksessa selvitimme kriittisen diskurssianalyysin keinoin, miten ja minkälaisia käsityksiä tekoälystä ja automaattisesta päätöksenteosta rakennettiin aineistossa, sekä miten Arviomuistiossa esitetty lainsäädäntöehdotelma suhteutui näihin käsityksiin. Tiukka rajaus ainoastaan lainsäädäntötyöryhmän jäsenten lausuntoihin antoi mahdollisuuden keskittyä nimenomaan sellaisiin diskursseihin, joilla katsoimme aidosti olevan vaikutusta lainsäädännön muodostumiseen. Aineistosta havaitsimme viisi toisistaan selvästi erottuvaa tekoälydiskurssia. Viidestä diskurssista neljä oli positiivisia. Niissä tekoälystä rakennettiin kuvaa 1) kehittäjänä, 2) uuden mahdollistajana, 3) ihmisen kaltaisena ja 4) vääjäämättömänä. Ainoastaan viides diskurssi oli jokseenkin pessimistinen ja tekoälyn suhteen varovainen. Siinä tekoäly nähtiin 5) riskinä.

Optimistisia diskursseja luovat ja ylläpitävät lausumat kuvasivat tekoälyä ja automaatiota joko kyseenalaistamattoman positiivisesti tai vähintään neutraalisti, ja diskursseista rakentui kuva vahvasta uskosta siihen, että tekoälyllä olisi runsaasti potentiaalia palveluiden tehostamiseen, kehittämiseen ja uudistamiseen. Ensimmäisessä diskurssissa tekoälystä muodostui kuva *kehittäjänä*. Diskurssissa painottuivat näkemykset siitä, että tekoälyn ja automaation avulla voitaisiin tehdä olemassa olevista palveluista tehokkaampia, minimoida virheet ja vapauttaa ihmisiä haastavampiin tehtäviin. Toisessa diskurssissa tekoäly kuvastui *uuden mahdollistajana*. Diskurssissa korostettiin, että tekoälyn avulla voitaisiin luoda uusia palveluita ja sitä kautta parantaa toiminnan laatua. Kolmannessa diskurssissa tekoälystä maalattiin kuva *ihmisen kaltaisena* päätöksentekijänä. Automaattisen päätöksenteon ei diskurssissa nähty juurikaan eroavan inhimillisestä päätöksenteosta, ja sen mahdollisesti luomat riskit samastettiin ihmisvetoisen päätöksenteon riskeihin. Jo olemassa olevien käytänteiden argumentoitiin riittävän suojaamaan myös tekoälyn mukanaan tuomilta riskeiltä. Neljännessä diskurssissa tekoälyn kehitys ja laajeneva käyttö esitettiin *vääjäämättömänä.* Tekoäly kuvattiin väistämättömänä osana tulevaisuuden yhteiskuntaa, mikä auttoi perustelemaan näkemystä siitä, että sen käyttöä ei tulisi lainsäädännöllä rajoittaa liikaa; päinvastoin, diskurssin mukaan kehitykseen tulisi varautua ja lainsäädännöstä luoda joustavaa ja mahdollistavaa.

Optimistisissa diskursseissa esitetyssä riskipuheessa tekoälyä itsessään ei nähty riskinä, vaan riskeinä esiin nousivat palveluiden hitaus ja saatavuusongelmat, joihin tekoäly ja automaatio voisivat tarjota ratkaisuja. Diskurssien voimakkuutta selittänee modernin yhteiskunnan jatkuvasti kiihtyvästä vauhdista seuraavat valtionhallinnon

[103] Ks. tarkka aineisto Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707. 4.

ongelmat[104], kapitalistinen tehokkuusajattelu sekä jo Suomen tekoälystrategiassa[105] esitetty pyrkimys toimia suunnannäyttäjänä (eettisen) tekoälyn kehityksessä ja käyttöönotossa[106]. Huomasimme, että samalla diskurssit kuitenkin etääntyivät sosiaalidemokraattisissa järjestelmissä usein korostetuista sosiaalisen oikeudenmukaisuuden teemoista. Keveää sääntelyä ja uusia mahdollisuuksia epäröimättä ihannoivat diskurssit kiinnittyivät oikeistolaiselle markkinataloudelle tyypillisiin arvoihin. Siinä missä Suomi on usein vaikuttanut olevan vastustuskykyinen (uus)liberalistisille ihanteille[107], aineistosta esiin nousseet optimistiset diskurssit toistivat tehokkuusajattelun eetosta. Optimistisissa diskursseissa myös häivytettiin automaation riskipotentiaali, mikä osaltaan vaikuttanee niissä esiin nostettuihin näkemyksiin perustuslaillisten velvoitteiden tulkinnasta. Esimerkiksi legaliteettiperiaatteen tiukka tulkinta saatiin optimistisissa diskursseissa vaikuttamaan lähinnä kehitystä hidastavalta seikalta. Sen velvoittavuutta häivytettiin yhtäältä rinnastamalla automaattinen päätöksenteko ihmislähtöiseen päätöksentekoon ja toisaalta kuvaamalla se puhtaasti mekaanisena ihmisen tahdon toteuttamisena, jolloin erityinen automaation sääntely saatiin näyttämään tarpeettomalta.

Missään diskursseissa yhteiskunnalliset haitat eivät juuri saaneet huomiota. Ongelmien monimutkaisuutta tai niiden laajempaa merkitystä käsiteltiin hyvin rajatusti. Viides diskurssi, jossa tekoäly käsitteellistettiin vahvasti riskinä, nousi aineistosta esille optimistisiin diskursseihin verrattuna heikosti. Riskidiskurssissa tekoäly kuvattiin vaikeasti hallittavaksi, ennalta-arvaamattomaksi ja läpinäkymättömäksi. Vaikka sinänsä diskurssissa maalattiin kuva tekoälystä riskialttiina teknologiana, riskien arviointi jäi lopulta riskidiskurssissakin kapea-alaiseksi. Riskien kuvattiin kohdistuvan perus- ja ihmisoikeuksiin ja hyvän hallinnon vaatimusten toteutusmahdollisuuksiin, eikä esimerkiksi laajempia yhteiskunnallisia vaikutuksia tuotu esille.

Analyysin tuloksena esitimme, että suhtautuminen tekoälyyn oli lainsäädäntöä valmistelevan työryhmän sisällä merkittävän optimistista. Optimistisuuden taustalla näyttivät vaikuttavan olemassa olevat valtionhallinnon ongelmat, joihin tekoälystä toivottiin nopeaa ja helppoa ratkaisua. Ratkaisujen hakemisen tarve näyttäisi kiinnittyvän tarpeeseen tarjota yhteiskunnassa tarpeellisia, lakisääteisiä palveluita riittävän

104 Ks. lisää Rosa, H. (2013). *Social Acceleration*. Columbia University Press. Luku 11.

105 Työ- ja elinkeinoministeriö. (2019). Edelläkävijänä tekoälyaikaan: Tekoälyohjelman loppuraportti. Työ- ja elinkeinoministeriön julkaisuja 2019:23.

106 Esimerkiksi Rosa argumentoi: ”[...] if and so long as politics wants to maintain its claim to regulate the parameters of economic and technological development, it must *either* adapt itself to the accelerated rate of innovation in the relevant social spheres and become, as it were, a “motorized legislator” (Carl Schmitt) *or* decisively intervene in their developmental autonomy and thereby repeal the principle of functional differentiation in favor of a renewed political dominance.” Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 263.

107 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

nopeasti. Aihepiiriä on tutkinut Rosa, joka argumentoi, että modernia yhteiskuntaa leimaa tarve jatkuvasti kiihtyvään muutokseen, joka kuitenkin yhteiskunnan eri osa-alueilla väistämättä tapahtuu jossain määrin eri tahdissa. Yhden yhteiskunnan osa-alueen muutosvauhdin kiihtyessä muiden osa-alueiden on reagoitava kiihdyttämällä vauhtia saman tahdin saavuttamiseksi, jotta eritahdistumisesta juontuvat haitat pysyvät mahdollisimman vähäisinä[108]. Jokseenkin samanlaista argumentaatiota pystyi havaitsemaan tekoälyyn optimistisesti suhtautuvista diskursseista: hallinnollisten instituutioiden olisi välttämätöntä tehostaa toimintaansa, jotta lisääntyvään palveluidentarpeeseen pystyttäisiin vastaamaan. Vaikeus tarjota lakisääteisiä palveluita perustuslain vaatimalla tavalla ilman aiheetonta viivytystä näkyikin optimistissa diskursseissa ensisijaisena ongelmana, ja tekoälytyökalut ratkaisuna siihen. Yhteiskunnan kiihtyvä vauhti vaikuttaisi siis lisäävän painetta hallinnollisten prosessien tehostamiseen ja samalla perustelevan optimistista suhtautumista tekoälytyökalujen sallimiseen: kun ongelmana nähdään byrokratian hitaus itsessään, sen ratkaiseminen on ensisijaista. Jos tekoäly ja automaatio nähdään olemassa olevien ongelmien ratkaisuna, niihin potentiaalisesti liittyvät, tulevaisuudessa mahdollisesti realisoituvat riskit jäävät toissijaisiksi tai jopa täysin vaille huomiota.

Toisaalta tunnistimme myös riskidiskurssista kiinnekohtia yhteiskunnan eri osa-alueiden eritahdistumiseen ja eri tahdissa kiihtyvän muutoksen mukanaan tuomiin ongelmiin. Optimistisissa diskursseissa keskityttiin ongelmiin, joita on jo syntynyt siksi, että yhteiskunnan vauhti on kiihtynyt, eivätkä hallinnolliset toimijat ole pystyneet vastaamaan lisääntyneisiin vaatimuksiin. Sen sijaan riskidiskurssissa nostettiin esille esimerkiksi automaattisten päätöksentekojärjestelmien seurauksena mahdollisesti lisääntyvä tarve muutoksenhaulle, jonka seurauksena muutoksenhakuinstanssien työtaakka voisi kasvaa kestämättömällä tavalla. Optimistisissa diskursseissa siis katsottiin, että tekoäly voisi auttaa ratkaisemaan olemassa olevia eritahdistumisen ongelmia, kun taas pessimistisessä riskidiskurssissa tekoälyn epäiltiin lisäävän eritahdistumista ja siitä seuraavia ongelmia tulevaisuudessa: virheellisten päätösten mahdollisesti kasvava määrä voisi siirtää eritahdistumisen ja sitä seuraavien ongelmien painopisteen päätöksentekovaiheesta oikaisu- ja muutoksenhakuvaiheisiin.

2.4 Osatutkimus 3: Between Algorithmic and Analogue Harms

Kolmannessa osatutkimuksessa jatkoimme Hanna Malikin kanssa siitä, mihin toisessa osatutkimuksessa olimme päätyneet. Artikkelissa *Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services* tarkaste-

[108] Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

limme algoritmisten teknologioiden, etenkin automaattisen päätöksenteon, ominaisuuksia ja mahdollisuuksia yhteiskunnallisten haittojen kontekstissa. Otimme esimerkiksi Maahanmuuttoviraston (Migri) pyrkimyksen automatisoida toimintaansa, ja sen kautta selvitimme, (1) miten Suomen tapa suhtautua automaattiseen päätöksentekoon sopi yhteen niiden perinteiden kanssa, jotka ovat nostaneet Suomen harm reduction regimes[109] -jaottelussa yhdeksi vähiten haitallisista valtioista ja (2) miten erilaisten haittojen yhteydet tunnistettiin ja huomioitiin tietyissä poliittisissa ja lainsäädännöllisissä dokumenteissa.

Osatutkimuksen tematiikka kiinnittyy yhteiskunnallisten haittojen tutkimusperinteeseen, johon tukeutuen tarkastelimme yhteiskunnallisten haittojen muodostumista ja ilmenemistä maahanmuutto- ja tekoälypolitiikan risteyskohdassa. Jaoimme artikkelissa haitat niin sanottuihin algoritmisiin ja analogisiin haittoihin: *algoritmisilla haitoilla* viittaamme ensimmäisen osatutkimuksen[110] tapaan niihin haittoihin, joita algoritmisten teknologioiden käytöstä syntyy tai voi syntyä, kun taas *analogisilla haitoilla* viittaamme yhteiskunnallisiin haittoihin, joiden syntymiseen algoritmit eivät perinteisesti ole olleet vaikuttamassa.

Artikkelin keskiössä vaikuttivat kaksi hallituksen esitystä, joiden tarkoituksena oli lisätä Maahanmuuttoviraston automaatiomahdollisuuksia sektorikohtaisella sääntelyllä: HE 224/2018, joka raukesi hallituksen vaihtumiseen, ja HE 18/2019[111]. Jälkimmäiseen antamassaan lausunnossa perustuslakivaliokunta esitti, että ennen kuin automaatiota voi julkishallinnollisessa päätöksenteossa hyödyntää, tulee selvittää mahdollinen tarve yleislaintasoiselle sääntelylle[112]. Lausunto käytännössä kielsi automaattisen päätöksenteon julkishallinnon kontekstissa mahdollisen yleislain säätämiseen asti. Kielto sysäsi käyntiin mielenkiintoisen tapahtumakulun. Yhtäältä se esti tehokkaasti potentiaalisten algoritmisten haittojen syntymisen. Samaan aikaan se kuitenkin esti myös päätöksentekoprosessien tehostamisen automaation avulla. Hitaista prosesseista ja viivästyneistä päätöksistä syntyvät, kiistatta olemassa olevat analogiset haitat jäivät siis vaille ratkaisua.

Maahanmuuttoviraston päätöksenteko oli ruuhkautunut jo vuoden 2015 niin sanotun pakolaiskriisin aikaan, minkä seurauksena vielä vuoden 2020 toimintakerto-

109 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

110 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. International Journal for Crime, Justice and Social Democracy, 11(1), 182–195.

111 HE 224/2018 ja HE 18/2019. Hallituksen esitys eduskunnalle laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

112 PeVL 7/2019. Perustusvaliokunnan lausunto hallituksen esityksestä laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

muksissaan sekä oikeuskansleri[113] että eduskunnan oikeusasiamies[114] katsoivat Migrin rikkovan toistuvasti vaatimuksia kohtuullisista käsittelyajoista. Tutkimusten perusteella on selvää, että pitkäksi venyneet käsittelyajat lisäävät haittoja, joita turvapaikanhakijat kohtaavat, ja joita tiukka maahanmuuttopolitiikka jo itsessään aiheuttaa maahan saapuville[115]. Epävarma tilanne, jonka kestosta ei ole tietoa, aiheuttaa ajallisia haittoja (*temporal harms*)[116], esimerkiksi mielenterveydellisiä ongelmia[117].

Analogisiin haittoihin oli pyritty löytämään erilaisia analogisia ratkaisuja. Migrin resursseja oli kasvatettu ja uusia työntekijöitä palkattu purkamaan turvapaikkahallinnon kriisiytynyttä tilannetta[118]. Silloisen poliittisen ilmapiirin mukaisesti lainsäädäntöä myös kiristettiin turvapaikanhakijamäärän vähentämiseksi ja turvapaikanhakijoiden oikeuksien rajoittamiseksi[119]. Ratkaisuyritykset loivat kuitenkin uusia ongelmia. Esimerkiksi uusien hakemuskäsittelijöiden palkkaaminen ja kouluttaminen hallinnollisen kriisin ratkaisemiseksi toivat mukanaan vakioratkaisupohjien hyödyntämisen[120], mikä vaaransi myös turvapaikanhakijoiden perus- ja ihmisoikeuksien toteutumisen[121].

Hallituksen esitysten tavoitteena oli mahdollistaa tekoälyavusteinen päätöksenteko, josta toivottiin kustannustehokasta ratkaisua hitaisiin prosesseihin. Sektorikohtaisella sääntelyllä pyrittiin mahdollistamaan automaattiset ratkaisumenetelmät Maahanmuuttovirastossa. Perustuslakivaliokunta katsoi kuitenkin, että automaation mahdollistamiseen pyrkivissä hallituksen esityksissä ei kyetty riittävällä tavalla huomioimaan perustuslaillisia vaatimuksia, kuten hyvän hallinnon periaatteita, tai EU:n yleisen tietosuoja-asetuksen vaatimuksia. Valiokunta totesi, että ehdotettu sääntely oli liian epätarkkaa ja jätti epäselväksi, miten perustuslailliset vaatimukset todella

113 Oikeuskanslerin virasto. (2020) Valtioneuvoston oikeuskanslerin kertomus vuodelta 2019. OKK 12/2020. 150–152.

114 Eduskunnan oikeusasiamies. (2020) Eduskunnan oikeusasiamiehen kertomus vuodelta 2019. OAK 15/2020. 224.

115 Canning, V., & Tombs, S. (2021). From social harm to zemiology: A critical introduction. Routledge. 83.

116 Canning, V. (2019). Abject asylum: Degradation and the deliberate infliction of harm against refugees in Britain. *Justice, Power and Resistance*, *3*(1), 37–60.

117 Soliman, F. (2021). States of exception, human rights, and social harm: Towards a border zemiology. *Theoretical Criminology*, *25*(2), 228–248.

118 Yle. (2017). Report: Immigration Service circulates model negative asylum decisions for ‘assembly line’ use. (4.5.2017).

119 Hallituksen turvapaikkapoliittinen toimenpideohjelma. (2015.)

120 Yle. (2017). Report: Immigration Service circulates model negative asylum decisions for ‘assembly line’ use. (4.5.2017).

121 Pirjatanniemi, E., Lilja, I., Helminen, M., Vainio, K., Lepola, O., & Alvesalo-Kuusi, A. (2021). Ulkomaalaislain ja sen soveltamiskäytännön muutosten yhteisvaikutukset kansainvälistä suojelua hakeneiden ja saaneiden asemaan. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 2021:10. 229.

täytettäisiin. Tämä merkitsi julkishallinnon automaatioprojektien keskeytymistä siihen asti, että tarve yleislaintasoiselle automaation sääntelylle olisi kartoitettu ja tarvittaessa laki säädetty.

Perustuslakivaliokunnan näkökanta muodosti mielenkiintoisen ristiriidan tekoälypoliittisissa dokumenteissa esitettyjen, tekoälyn mahdollisuuksia painottavien tavoitteiden kanssa. Tekoälypolitiikassa korostettiin tekoälyn tarjoamia mahdollisuuksia ja tarvetta mahdollistaa tekoälyn laaja-alainen hyödyntäminen. Varhaisessa tekoälypolitiikassa, kuten esimerkiksi Tekoälyaika-projektin puitteissa luodussa Suomen tekoälystrategiassa[122], korostettiin ihmiskeskeisyyttä tärkeänä osana tekoälyn käyttöönottoa. Uudemmissa tekoälypoliittisissa dokumenteissa ihmiskeskeisyyden merkitys kuitenkin vähentyi, ja esimerkiksi vuonna 2020 alkaneessa Tekoäly 4.0-projektissa keskityttiin vahvasti yritysten kilpailukyvyn edistämiseen ja valtion taloudelliseen kasvuun[123]. Osatutkimuksessa nostimmekin esiin huolen siitä, että poliittisten äänenpainojen muutos ihmiskeskeisestä yrityskeskeiseksi vaikuttaa vahvistavan uusliberalistisia näkemyksiä Suomessa. Pidemmällä aikavälillä se voi muuttaa Suomen yhteiskuntajärjestystä tavalla, joka vaikeuttaa yhteiskunnallisten haittojen hallintaa[124].

Tutkimme aineistostamme myös, miten eri poliittisilla ja oikeudellisilla sektoreilla ymmärretään tekoäly ja sen vaikutukset, mahdolliset yhteiskunnalliset haitat ja etenkin tekoälyn vaikutukset yhteiskunnallisiin haittoihin. Analyysin perusteella argumentoimme, että siinä missä maahanmuuttopolitiikassa yhteiskunnallisia haittoja on ymmärretty ja pohdittu melko laaja-alaisesti, ymmärrys tekoälystä on jäänyt hyvin kapeaksi tai jopa olemattomaksi, mikä osaltaan voi estää teknologisten ratkaisujen kehittämistä ja käyttöönottoa. Tekoälypolitiikassa potentiaalisista haitoista on sen sijaan käyty keskustelua vain vähän. Siinä missä maahanmuuttopolitiikassa haittojen kärsijänä on kuvattu yksilö ja/tai yhteiskunta, tekoälypolitiikassa on keskitytty yrityksiin kohdistuviin taloudellisiin haittoihin, joita kallis ihmistyövoima tuottaa.

Oikeudellisissa teksteissä riskejä ja haittoja on arvioitu lähinnä oikeudellisten velvoitteiden näkökulmasta. Hallituksen esityksissä huomiota annettiin erityisesti tehottomille prosesseille, joita automaation avulla olisi mahdollista kehittää, kun taas perustuslakivaliokunta täysin päinvastoin keskittyi riskeihin, joita automaation avulla toteutettu tehostaminen luo hallinnollisten prosessien perustuslainmukaisuudelle. Yksilöihin tai yhteisöihin kohdistuvat riskit ja haitat pelkistyivät molemmissa

122 Työ- ja elinkeinoministeriö. (2019). Edelläkävijänä tekoälyaikaan: Tekoälyohjelman loppuraportti. Työ- ja elinkeinoministeriön julkaisuja 2019:23.

123 ”Tekoäly 4.0 -ohjelman visio ’Suomesta voittaja kaksoissiirtymässä’ kytkeytyy laajemmin tavoitteisiin kilpailukykyisestä, ilmastoneutraalista ja digitalisoituneesta teollisuudesta.” Työ- ja elinkeinoministeriö. (2022). Tekoäly 4.0 -ohjelman loppuraportti. Työ- ja elinkeinoministeriön julkaisuja 2022:60. 7.

124 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

tapauksissa potentiaalisiksi seurauksiksi siitä, että lakia ei noudateta, mitä ei nimenomaisesti kuitenkaan nostettu esille.

Analyysin perusteella on ilmeistä, että artikkelin kirjoittamisen aikana eri sektorit eivät tehneet riittävästi yhteistyötä, jotta poliittisten toimijoiden olisi ollut mahdollista hahmottaa kokonaiskuva tekoälyn hyödyntämisen mahdollisuuksista ja vaikutuksista. Siiloutunut poliittinen kenttä olisi saatava yhdistymään, jotta tekoälyjärjestelmien käytöstä seuraavia potentiaalisia haittoja ja toisaalta hyötyjä voitaisiin ymmärtää kokonaisuutena. Vaikuttaa siltä, että tekoälypolitiikassa oli varsin heikosti ymmärretty sekä käytännön todellisuutta esimerkiksi maahanmuuttohallinnossa että perustuslaillisia velvoitteita. Toisaalta maahanmuuttopoliittisessa keskustelussa käsiteltiin tekoälyn avaamia mahdollisuuksia hyvin heikosti. Toisiinsa liittyvien aihepiirien käsittely erillisinä oli väistämättä johtanut siihen, että huomiotta jäi potentiaalisia ja myös jo olemassa olevia ongelmia. Tekoälyn tarjoamat mahdollisuudet olivat poliittisesta innostuksesta huolimatta jääneet yhteiskunnan monilla sektoreilla tunnistamatta. Artikkelissa argumentoimmekin, että jo olemassa olevien analogisten ja potentiaalisten algoritmisten haittojen ratkaisemiseksi tulisi holistista näkökulmaa ja eri hallinnonalojen välistä yhteistyötä lisätä.

3 Tekoälyjärjestelmiin liittyvät haitat

Tässä luvussa pääosassa ovat haitat, joita algoritmisten teknologioiden hyödyntämisestä voi aiheutua tai on jo aiheutunut. Algoritmien merkitys yhteiskunnissa ja ihmisten elämissä on kasvanut, ja samalla myös algoritmisten teknologioiden vaikutukset ovat lisääntyneet, kumuloituneet ja kerrostuneet niin mikro-, meso- kuin makrotason vuorovaikutussuhteissa. Kun tällaisten läpitunkevien ja häiritsevien (*disruptive*)[125] teknologioiden toiminta on usein vaikeasti havaittavaa, myös käsitys niihin liittyvistä haitoista ja haittojen syntymekaniikasta saattaa hämärtyä[126]. Tällä hetkellä näyttää siltä, että kehitys jatkaa samaan suuntaan: algoritmien ja algoritmisten teknologioiden vaikutukset yhteiskunnassa lisääntyvät entisestään, ja paine algoritmisaation kiihdyttämiseen kasvaa, kuten osatutkimusten perusteella on nähtävissä. Mahdollisesti seuraavien haittojen tutkiminen käy yhä merkityksellisemmäksi – ja toisaalta myös haastavammaksi.

Haitat, jotka etääntyvät välittömästi teknologian toiminnasta seuraavista yksilötason haitoista, syntyvät yhä monimutkaisempien vuorovaikutussuhteiden ja kausaaliketjujen seurauksena. Yksittäisten teknologioiden vaikutuksia on yhä vaikeampi erottaa erityisesti, kun aikajänne pitenee ja vaikutusalue laajenee. Usein käyttäjien on myös vaikea saada tietoa paitsi hyödynnettyjen algoritmien toiminnan logiikasta myös ylipäätään niiden olemassaolosta[127]. Esimerkiksi Smuha puhuu tietämyskuilusta (*knowledge gap*): on vaikea tunnistaa haittoja, jos ne syntyvät läpinäkymättömien, mahdollisesti näkymättömissä toimivien algoritmien toiminnan seurauk-

125 Suomenkielinen termi on jokseenkin kömpelö. Englanninkielisellä termillä *disruptive technologies* viitataan teknologioihin, joilla on kyky muovata ihmisten, yritysten ja instituutioiden toimintaa, ja jotka ominaisuuksiensa ansiosta syrjäyttävät totuttuja tapoja ja toiminnan muotoja. Tämän voi argumentoida ”häiritsevän” aiemmin vallinnutta järjestystä. Ks. esimerkiksi Latzer, M. (2009). Information and communication technology innovations: radical and disruptive? *New Media & Society*, *11*(4), 599–619.

126 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

127 Elliott, A. (2019). The culture of AI: Everyday life and the digital revolution. Routledge.

sena[128]. Toisaalta tunnistettujenkin haittojen yhteyttä tiettyyn teknologiaan voi olla mahdotonta osoittaa.

Vaikka algoritmisten teknologioiden vaikutuksia voi olla vaikea havaita, ne kuitenkin muovaavat ihmisiä ja yhteiskuntia ennennäkemättömällä tavalla. Nykypäivän moderneissa yhteiskunnissa lähes kaikki inhimillinen toiminta sitoutuu tavalla tai toisella teknosfääriin, teknologioiden tarjoamaan ympäristöön[129]. Tämä on yksi ilmentymä Rosan kuvaamasta yhteiskunnallisesta kiihtymisestä[130]. Jos on saavutettava yhä enemmän yhä lyhyemmässä ajassa ja yhä vaativammilla standardeilla, uudet työkalut, kuten älykkäät teknologiat, muodostavat pelastusrenkaan, jonka avulla voi selvitä yhä tiiviimmästä ja tempoltaan kiihtyvästä elämästä. Teknosfääriin, kuten muihinkin kiihtymisen ilmentymiin sopeutuminen muodostuu näkymättömäksi pakoksi, jonka välttämiseksi on olemassa hyvin vähän mahdollisuuksia[131]. *Välttämättömyys* sopeutua, *synkronoitua* ympäristön tahtiin, lisää entisestään tekoälyteknologioiden houkuttelevuutta läpi yhteiskunnan osa-alueiden. Näin ollen tekoälyteknologiat pääsevät vaikuttamaan jatkuvasti syvällisemmin ja laajemmin niin inhimilliseen toimintaan, kulttuuriin ja sosiaalisiin normeihin kuin yhteiskunnan instituutioihin ja perustavanlaatuisiin rakenteisiinkin. Tämä muuttaa väistämättä yhteiskunnallista vallanjakoa ja kapitalistisen yhteiskunnan prosesseja[132]. Ensimmäisessä osatutkimuksessa kollegoideni kanssa argumentoimmekin, että nykyaikana voisi olla tarpeellista puhua sosioekono*teknisistä* järjestelmistä ja rakenteista. Yksittäisten teknologioiden tarkastelun ohella tulisi tarkastella teknologioita järjestelmätason ilmiöinä, jotka muovaavat yhteiskunnan rakenteita, instituutioita, ihmisiä ja näiden vuorovaikutusta[133].

Yhteiskunnan ja sen toimijoiden välistä vuorovaikutusta teoretisoi yhteiskuntatieteellinen rakenteistumisteoria[134], jonka keskeisenä argumenttina on, että siinä missä inhimilliset toimijat muokkaavat yhteiskunnan rakenteita, nämä rakenteet muovaavat myös inhimillisiä toimijoita. Toisin sanoen yhteiskunnallinen todellisuus *rakenteistuu* toimijoiden ja rakenteiden vuorovaikutussuhteessa. Muutokset voivat johtaa paitsi kehitykseen kohti parempaa ja toimivampaa myös erilaisiin haitallisiin lopputulemiin. Vuorovaikutussuhdetta toimijan ja rakenteiden välillä ei tulisikaan tarkastella yksisuuntaisena vaan molempiin suuntiin vaikuttavana ilmiönä. Myös

128 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3). 9.

129 Ruuska, T., & Heikkurinen, P. (2023). Kokoava ote teknologiaan Marxiin ja Heideggeriin pohjautuen. *Tutkimus & kritiikki*, *3*(1), 68–87.

130 Rosa, H. (2013). *Social acceleration.* Columbia University Press.

131 Rosa, H. (2013). *Social acceleration.* Columbia University Press.

132 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.

133 Myös Ruuska, T., & Heikkurinen, P. (2023). Kokoava ote teknologiaan Marxiin ja Heideggeriin pohjautuen. *Tutkimus & kritiikki*, *3*(1), 68–87.

134 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

epätoivotut seuraukset, haitat, voivat siis kohdistua paitsi yksilöön, myös rakenteeseen. Kun sekä yksilöiden toiminta että yhteiskunnan rakenne ovat yhä tiukemmin sidottuja teknosfääriin, teknologiat ohjaavat myös rakenteistumista yhä voimakkaammin. Tämä on merkityksellistä: jos sekä toiminta että rakenteet toimivat teknosfäärin tarjoamien mahdollisuuksien kautta, vuorovaikutus pelkistyy teknosfäärin tarjoamien raamien sisälle. Tämä tarkoittaa, että algoritmiset teknologiat eivät niinkään monipuolista yhteiskunnallista todellisuutta ja lisää toiminnan mahdollisuuksia, vaan niiden hyödyntäminen itse asiassa kaventaa niin yhteiskunnallista kuin inhimillistä monimuotoisuutta.[135]

Teknologioiden, ihmisten ja yhteiskunnallisten ilmiöiden vuorovaikutusta on tutkittu pitkään[136]. Algoritmiset teknologiat ovat tässä vuorovaikutussuhteessa vielä jokseenkin tuore tulokas. ja niihin liittyviä haittoja on vasta alettu ymmärtää. Länsimaisissa yhteiskunnissa on jo jokseenkin laajasti tunnistettu sellaiset yksittäisiin teknologioihin, niiden ominaisuuksiin ja toimintalogiikkaan kiinnittyneet haitat, jotka ilmenevät välittömästi teknologioiden toiminnan seurauksena. Poliittisessa ja oikeudellisessa keskustelussa on nostettu esiin esimerkiksi vinoumat (bias), läpinäkymättömän ja/tai selittämättömän toimintalogiikan riskit (opacity, black box problem), yksityisyydensuoja (privacy) ja toisaalta erilaiset vihamieliseen toimintaan liittyvät turvallisuusuhkat[137]. Tällaiset haitat ovat tietenkin todellisia ja niihin puuttuminen välttämätöntä. Ratkaisuiksi on usein tarjottu selitettävyyden lisäämistä (explainable AI), läpinäkyvyyttä (tranparency), laadukkaampaa opetusdataa, ja panostamista teknologian luotettavuuteen (reliability/trustworthiness). Nämä voivatkin vähentää algoritmisten teknologioiden toiminnasta välittömästi seuraavia haittoja. Välittömien, yksittäisiin teknologioihin kiinnittyvien vaikutusten lisäksi algoritmisilla järjestelmillä on kuitenkin myös potentiaalia muuttaa yhteiskuntia radikaalisti. Algoritmisten teknologioiden radikaali muutosvoima näyttäisikin kiinnittyvän siihen, että tekoälyteknologioiden vaikutukset eivät rajaudu yksittäisen teknologian suoriin, ajassa ja tilassa välittömästi teknologiaan kiinnittyviin vaikutuksiin.

Läpi modernin yhteiskunnallisen todellisuuden tapahtumat seuraavat toisiaan yhä kiihtyvällä vauhdilla. Rosa kuvaa tätä ilmiötä nykyhetken *tiivistymisenä:* mennyt ja tuleva kirjoitetaan uudestaan yhä nopeammin, jolloin nykyhetken merkitys

135 Malik, H. M., Lepinkäinen, N., Alvesalo-Kuusi, A., & Viljanen, M. (2022). Social harms in an algorithmic context. *Justice, Power and Resistance*, *5*(3). 194.

136 Esimerkiksi Ellul, J. (1964). *The technological society*. New York: Vintage;
Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy*, *25*(6), 865–899;
Floridi, L. (2014). The fourth revolution: how the infosphere is reshaping human reality. Oxford: Oxford University Press.

137 Ks. Hine, E., & Floridi, L. (2023). The Blueprint for an AI Bill of Rights: in search of enaction, at risk of inaction. *Minds and Machines*, 1–8.

todellisuuden kuvaajana heikkenee[138]. Yhtäältä tästä seuraa, että algoritmisten teknologioiden vaikutukset leviävät ja kertautuvat ajallisesti tiivistyvässä teknosfäärissä yhä nopeammin. Vuorovaikutussuhteet käyvät yllättävämmiksi ja vaikeammin hahmotettaviksi. Toisaalta algoritmiset teknologiat tarjoavat keinon *sopeutua* ajallisesti tiivistyneeseen yhteiskuntaan[139]. Tämä myös tarkoittaa, että moderni maailma muotoutuu yhä laajemmin algoritmisten teknologioiden ympärille: on käytännössä mahdotonta määritellä rajaa on- ja offline-elämän välille[140]. Näin ollen yksittäisenkin tekoälyteknologian haitalliset vaikutukset voivat nousta esiin vasta ajan myötä, kun riittävän monien teknologioiden kumuloituvat vaikutukset ylittävät kriittisen pisteen, jonka seurauksena haittoja alkaa muodostua.

Tekoälyteknologioiden vaikutus kattaa koko inhimillisen elämän piirin. Se kohdistuu niin yksilöiden arkielämän toimintaan, sosiaalisiin suhteisiin ja tiedonsaantiin[141] kuin myös instituutioiden toimintaan ja työelämään[142], demokratiaan[143] ja sosiaalisen oikeudenmukaisuuden toteutumiseen[144]. Vaikutukset ovat valtavia. On jokseenkin väistämätöntä, että ne kohdistuvat myös ihmisten ajatteluun, arvoihin ja toimintaan: ihmisen toiminta ja ajattelu ovat väistämättä sidoksissa ympäristön tarjoamiin reunaehtoihin. Tähän perustuen on jopa argumentoitu, että tekoälyteknologioiden käynnistämä transformaatio uhkaa perustavanlaatuisella tavalla ihmiskeskeisen historian, kulttuurin ja sivistyksen – jotka voidaan tulkita hyvän elämän välttämättömiksi rakennuspalikoiksi – olemassaoloa, ja vaarana on uuden konekeskeisen maailmanjärjestyksen muotoutuminen[145]. Mikäli kehitys todella vie tähän suuntaan, antroposeenistä, ihmisten ajasta, siirrytään automaseeniin[146], koneiden aikaan, jossa ihmisten ajattelua ja toimintaa hallitsevat älykkäät koneet[147]. Ilman ajattelun ja toiminnan vapautta mahdollisuus tavoitella hyvää elämää väistämättä rapautuu. Jos hai-

138 Ks. Rosa, H. (2013). *Social acceleration.* Columbia University Press. 77.
139 Rosa, H. (2013). *Social acceleration.* Columbia University Press. 156.
140 Willson, M. (2019). Algorithms (and the) everyday. Teoksessa *The Social Power of Algorithms* (pp. 137–150). Routledge.
141 Ibid.
142 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research.
143 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).
144 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.
145 Harari, Y. (2023). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. (28.4.2023). *The Economist.*
146 Flanagan, M. (2018). The rise of the "Automacene": How robots will define the next epoch in human history. (16.6.2018). *Salon*.
147 Jokseenkin vastaavassa merkityksessä on puhuttu myös ”algoritmoseenistä”. Ks. Lee, S. (2020). Our Short-Lived Anthropocene and the Coming Algorithmocene. (20.9.2020). *Medium*.

tat ymmärretään hyvää elämää uhkaaviksi tai estäviksi seikoiksi, tällainen uusi maailmanjärjestys voisi olla yksi mahdollinen haitallisten kehityskulkujen päätepiste.

Tulevaisuuden ennustaminen on luonnollisesti mahdotonta. Valistuneita näkemyksiä on kuitenkin mahdollista esittää. Jo nyt on nähtävillä, että tekoälyteknologiat muuttavat yhteiskuntia niiden rakenteellisia perustuksia myöden. Jo havaittavissa olevien muutosten ja niiden vaikutusten tunnistaminen ja arviointi on yhteiskunnallisesti ja oikeudellisesti ensiarvoisen tärkeää, jotta teknologioiden sääntelyä, kehitystä ja käyttöä voidaan suunnata siten, että teknologioiden käyttöönotto ei rapauta hyvän elämän edellytyksiä. Pyrkimys voi olla vaikeasti toteutettavissa: niin kauan kuin tahot, joilla on suurin mahdollisuus vaikuttaa teknologioiden kehitykseen ja sääntelyyn – eli etenkin suurvallat ja tietyt teknologiajätit – hyötyvät enemmän tekoälyteknologioiden laajasta tuotannosta ja kevyestä sääntelystä kuin niiden rajoittamisesta, haittoja vähentäville toimille ei välttämättä saada tarvittavaa kannatusta. Maailmantalous, jonka lainalaisuuksien mukaan myös teknologiateollisuus rakentuu, muovautuu pitkälti siis voimakkaimpien toimijoiden ehdoilla. Tämä luonnollisesti altistaa myös haitoille. Mahdollisten – ja jo toteutuneiden – haittojen vakavuuden esiin nostaminen kuitenkin toivoakseni lisää ymmärrystä riskeistä, joita hallitsemattomaan algoritmiseen transformaatioon liittyy, ja auttaa hahmottamaan, minkä takia tehokas, haittoja ennalta estävä sääntely olisi ihmisten hyvinvoinnin kannalta välttämätön.

Seuraavissa alaluvuissa käsittelen haittojen tutkimuksen lähtökohtia, haittojen typologioita ja erilaisia tapoja arvioida yhteiskunnallisia, erityisesti algoritmisiin teknologioihin kiinnittyviä haittoja. Niiden jälkeen siirryn arvioimaan algoritmisten teknologioiden aikaansaamia haittoja. Niiden analyysissä otan huomioon 1) haitan ilmenemisen muodon, 2) teknologian ja haitan välisen etäisyyden sekä 3) haitan kohteen. Pyrkimyksenäni ei ole kuvata kaikkia mahdollisia tekoälyteknologioihin liittyviä haittoja vaan tuoda ilmi haittojen moninaisuus ja algoritmisten teknologioiden vaikutusten laaja-alaisuus.

3.1 Algoritmisten haittojen arvioimisen lähtökohtia

Erilaiset teknologiat ovat historian saatossa muuttaneet laajasti yhteiskunnallista todellisuutta. Tekoälyteknologiat ovat jatkumoa tälle historialle. Tekoäly on kuitenkin sovellettavuudessaan omassa luokassaan: erilaisia algoritmisia teknologioita hyödynnetään läpi yhteiskunnan, ja niiden vaikutuksilta on lähes mahdoton välttyä.

Jos ymmärrämme algoritmiset teknologiat laajasti, niiden kriittinen tutkimus on kulkenut 2000-luvun alun ohjelmistotutkimuksesta (software studies) ensin data- ja algoritmitutkimukseen (data studies, algorithm studies) ja siitä 2020-luvun taitteen tekoälytutkimukseen (critical AI studies). Teknologiat ovat kehittyneet siis koko ajan monimutkaisemmiksi, ja teknologioihin keskittyvän tutkimuksen painotukset ovat muuttuneet kehityksen mukana.

Tieteen- ja teknologiantutkimus (science and technology studies, STS) on tuonut voimakkaasti mukaan näkemyksen siitä, että tieto – jonka osana teknologinen kehitys voidaan nähdä – ja yhteiskunta kehittyvät rinnakkain ja myös määrittävät toistensa kehitystä (*co-production*)[148].

Voidaan argumentoida, että teknologioiden monimutkaistuminen seuraa tai *on osa* tiedon lisääntymistä. Monimutkaistumisesta taas seuraa se, että teknologioita ja niiden toimintoja ymmärtää yhä harvempi, jolloin myös yhä suurempi osa teknologioiden vaikutuksista käy yhä vaikeammin arvioitaviksi. Tieto on nopeasti lisääntynyt ja pirstaloitunut. Samaan aikaan myös modernia yhteiskuntaa määrittävät pirstaleisuus, nopeatempoisuus ja vaikeaselkoisuus sekä lukuisten toimijoiden ja sosiaalisten prosessien verkostot[149]. Haittojen tutkimisen kannalta tällä on merkitystä. Nopea yhteiskunnallinen ja teknologinen muutos aiheuttavat valtavasti *hälyä (noise),* jonka läpi on vaikea erottaa ja tutkia yksittäisiä ilmiöitä – mikäli risteävien ja kerrostuvien vaikutusten verkostosta ylipäätään voidaan erottaa toisistaan riippumattomia, erillisiä ilmiöitä. Todennäköisesti tämä ei ole mahdollista. Tämän takia kriittinen katse on kohdistettava laajempiin vuorovaikutussuhteisiin, joiden osana algoritmiset teknologiat vaikuttavat. Mikäli näin ei tehdä, riskinä on, että teknologioihin kiinnittyvät haitat sumentuvat ja sulautuvat ihmisten silmissä osaksi neutraaleja, luonnollisia ja vaikeasti ymmärrettäviä kehityskulkuja, jotka ovat vaikutusmahdollisuuksien ulkopuolella[150].

Haittojen tutkimus on osin eronnut valtavirtakriminologiasta juuri siksi, että perinteisen kriminologisen tutkimuksen piirissä haitat, jotka eivät täytä rikoksen ominaispiirteitä, on usein tulkittu luonnollisiksi, väistämättömiksi ja tahattomiksi ilmiöiksi, joihin puuttuminen ei ole ollut mahdollista[151]. Tämä lienee vaikuttanut siihen, että ne ovat jääneet vaille kriminologien suurempaa huomiota. Yhteiskunnallisten haittojen tutkijat kuitenkin argumentoivat laajasti, että haittojen taustalla voidaan nähdä kapitalistiselle järjestelmälle ominaisia piirteitä: vallan ja varallisuuden epätasainen jakautuminen, kapitalistinen kilpailu ja elämän eri osa-alueiden kaupallistuminen[152]. Samat yhteiskuntajärjestelmän piirteet vaikuttavat algoritmisen transformaation ja siihen kiinnittyvien haittojen taustalla, kuten ensimmäisessä osatutkimuk-

148 Jasanoff, S. (2004). The idiom of co-production. Teoksessa *States of knowledge* (1–12). Routledge.

149 Ks. myös Esko, T., & Koulu, R. (2022). Rethinking research on social harms in an algorithmic context. *Justice, Power and Resistance*, *5*(3), 307–313.

150 Ks. myös Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. International Journal for Crime, Justice and Social Democracy, 11(1), 182–195.

151 Ks. lisää esimerkiksi Friedrichs, D. O. (2009). *Trusted criminals: White collar crime in contemporary society*. Cengage Learning; ja
Pemberton, S. A. (2015). *Harmful societies: Understanding social harm*. Policy Press. 3–5.

152 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

sessa argumentoidaan[153]. Algoritmisetkaan haitat, kuten muutkaan yhteiskunnalliset haitat, eivät siis ole tahattomia tai luonnollisia, vaan ne tuotetaan yhteiskunnallisissa prosesseissa. Näin ollen niihin on mahdollista vaikuttaa vaikuttamalla yhteiskunnallisiin mekanismeihin ja rakenteisiin, joiden seurauksena ne ilmenevät.

Kriminologit ovat tutkineet tekoälyn ja algoritmien aiheuttamia yhteiskunnallisia haittoja melko vähän. Vaikka teknologisen kehityksen vaikutuksia on tutkittu ainakin 1980-luvulta lähtien, on kriminologien ja oikeustieteilijöiden näkökulma painottunut haittoihin, joita teknologiat ovat aiheuttaneet rikollisen toiminnan osana[154]. Viime aikoina on tutkittu myös esimerkiksi data-analyysin mahdollisuuksia rikosten ennalta estämisessä ja automaation hyödyntämistä esimerkiksi oikeudenkäytössä[155]. Yhteistä näille tutkimuksille on, että algoritmiset teknologiat nähdään välineinä, jotka mahdollistavat rikosten tekemisen, ennalta estämisen tai selvittämisen aiempaa tehokkaammin. Tutkittava aihepiiri on siis painottunut melko vahvasti perinteiseen rikollisuuteen ja siihen liittyviin prosesseihin, ja tekoäly on siten vertautunut työkaluun tai aseeseen[156].

Viime aikoina on julkaistu tutkimuksia myös siitä, miten algoritmisaation aiheuttamat haitat kohdentuvat toisaalta vähemmistöihin ja marginalisoituihin[157], toisaalta laajemmin yhteiskuntaan[158]. Näille näkemyksille yhteistä on, että teknologioiden vaikutusten ei nähdä pelkistyvän niiden käytön välittömiin seurauksiin. Oma tulo-

153 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

154 Ks esimerkiksi Smith, G. J., Bennett Moses, L., & Chan, J. (2017). The challenges of doing criminology in the big data era: Towards a digital and data-driven approach. *The British journal of criminology*, *57*(2), 259–274;

Van der Wagen, W., & Pieters, W. (2015). From cybercrime to cyborg crime: Botnets as hybrid criminal actor-networks. *British journal of criminology*, *55*(3), 578–595;

Wall, D. (Ed.). (2001). *Crime and the Internet*. Routledge.

155 Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology*, *18*(5), 623–642;

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. *The British Journal of Criminology*, *59*(3), 674–692.

156 King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics*, *26*, 89–120.

157 Stratton, G., Powell, A., & Cameron, R. (2017). Crime and justice in digital society: Towards a 'digital criminology'? *International Journal for Crime, Justice and Social Democracy*, *6*(2);

Ellis, J. R. (2022). Blurred consent and redistributed privacy: owning LGBTQ identity in surveillance capitalism. In *Diversity in Criminology and Criminal Justice Studies,* 183–196. Emerald Publishing Limited.

158 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).

kulmani asettuu osaksi tätä yhteiskunnallista, valtavirtakriminologiasta irtautuvaa kriittistä perinnettä.

Kriittisen kriminologian tutkijat korostavat, että yhteiskunnan toimintaa ja sen muutoksia ohjaavat tahot, joilla on tosiasiallista valtaa. Näistä tahoista puhutaan termillä *the powerful*[159]. Termi viittaa esimerkiksi ja erityisesti suuryrityksiin, valtiollisiin vallanpitäjiin ja globaaliin, vaikutusvaltaiseen eliittiin eli tahoihin, jotka pystyvät toimillaan vaikuttamaan merkittävästi siihen, millaisessa yhteiskunnallisessa todellisuudessa elämme. Kun valtaapitävät pääsevät ohjaamaan muutosta, josta usein puhutaan myös optimistisesti kehityksenä, riskinä on, että se hyödyttää eniten juuri niitä tahoja, joilla lähtökohtaisestikin on jo parhaimmat edellytykset hyvälle elämälle[160]. Valtavalle osalle ihmisiä hyödyt eivät sen sijaan välity, vaan päin vastoin moni joutuu kärsimään yhteiskunnallisen muutoksen mukanaan tuomista erilaisista haitoista[161]. Teknologinen kehitys ei poikkea tästä logiikasta. Teknologiat paitsi tehostavat toiminnan tapoja ja luovat uusia mahdollisuuksia *joillekin*, mikä voi olla osaltaan tarpeellista ja toivottua, myös heikentävät *toisten* hyvinvoinnin edellytyksiä ja altistavat epätasaisesti jakautuville haitallisille seurauksille.

Algoritmeihin ja tekoälyyn liittyvät mahdollisuudet, riskit ja haitat määritellään akateemisessa tutkimuksessa monin eri tavoin. Aihepiiri voidaan kehystää kysymyksillä etiikasta[162], tietosuojasta ja erityyppisistä oikeudenloukkauksista[163]. Monet ovat keskittyneet selvittämään algoritmien osuutta valtion valvontakoneistossa[164] tai niiden vaikutusta kapitalistiseen järjestelmään[165], markkinoihin[166], työelämään[167], globaaliin

159 Rothe, D., & Kauzlarich, D. (2016). *Crimes of the powerful: An introduction*. Routledge.

160 Myös esimerkiksi Ruuska, T., & Heikkurinen, P. (2023). Kokoava ote teknologiaan Marxiin ja Heideggeriin pohjautuen. *Tutkimus & kritiikki*, *3*(1), 68–87.

161 Whyte, D. (2017). Crime as a social relation of power: Reframing the 'ideal victim'of corporate crimes. Teoksessa *Handbook of Victims and Victimology* (pp. 333–347). Routledge.

162 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679;
Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117.

163 Todolí-Signes, A. (2019). Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: the necessary collective governance of data protection. *Transfer: European Review of Labour and Research*, *25*(4), 465–481.

164 Brayne, S. (2020). Predict and surveil: Data, discretion, and the future of policing. Oxford University Press, USA.

165 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.

166 Snider, L. (2014). Interrogating the algorithm: Debt, derivatives and the social reconstruction of stock market trading. *Critical Sociology*, *40*(5), 747–761.

167 Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., ... & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539.

eriarvoisuuteen[168] tai oikeudenmukaisuuteen[169]. Epätasainen vallan jakautuminen kuitenkin väistämättä luo myös eriarvoisuutta ja ylläpitää joidenkin ihmisryhmien vaurautta huonommassa asemassa olevien ryhmien kustannuksella, jolloin eriarvoisuutta ruokkiva ja siihen kytkeytyvä algoritminen transformaatio vääjäämättä myös luo haittoja yhteiskunnallisten haittojen tutkimusperinteessä ymmärretyllä tavalla. Tällaisia haittoja algoritmiset järjestelmät vaikuttaisivat luovan yhtäältä joko järjestelmän epäonnistumisen kautta tai toisaalta sinänsä onnistuneesti toteutetun järjestelmän mahdollisesti yllättävänä seurauksena[170]. Lisäksi algoritmisten järjestelmien avulla voidaan tietoisesti pyrkiä luomaan erilaisia haitallisia vaikutuksia, jotka eivät kuitenkaan täytä minkään rikoksen tunnusmerkistöä. Algoritmisten teknologioiden aikaansaamat todellisuuden muutokset avaavat myös uusia mahdollisia tulevaisuuksia, joista kaikki eivät todennäköisesti ole toivottavia. Tällaisia epätoivottuja tulevaisuuksia voi seurata myös sinänsä välittömästi positiiviseksi arvioiduista muutoksista.

Algoritmisten teknologioiden ja haittojen suhteessa korostuu odottamattomuus ja yllätyksellisyys. Tämä ei kuitenkaan tarkoita, että algoritmisia haittoja olisi täysin mahdotonta ennakoida tai hallita. Sen sijaan se tarkoittaa sitä, että teknologioiden monimutkaistuessa on jatkuvasti vaikeampaa varmistua siitä, että teknologian suunnittelijalla on kattava ymmärrys teknologian tulevasta käyttöympäristöstä ja toisaalta teknologian käyttäjällä riittävä ymmärrys teknologian toiminnan logiikasta, käytön merkityksestä ja mahdollisista seurauksista. Haittapotentiaalia määrittävätkin paitsi teknologian toiminta itsessään, myös sen välittömän toimintaympäristön (tekniset ja muut) ominaisuudet, algoritmiseen teknologiaan ja sen toimintaympäristöön liittyvät kaupalliset intressit, teknologian hyödyntämisen tavat ja käyttökonteksti[171]. Mitä monimutkaisemmassa ympäristössä algoritminen järjestelmä toimii, sitä vaikeampi on kattavasti ennakoida yhtäältä, miten se vaikuttaa ympäristöönsä, ja toisaalta, miten ympäristö vaikuttaa sen toimintaan tai käyttöön.

Kuten aiemmin olen esittänyt, on jokseenkin selvää, että algoritmisilla teknologioilla on valtava potentiaali lisätä eriarvoisuutta, voimistaa olemassa olevia hai-

168 Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197.

169 Mann, M. (2020). Technological politics of automated welfare surveillance: Social (and data) justice through critical qualitative inquiry. *Global Perspectives*, *1*(1).

170 Näin, mikäli rajataan pois rikosoikeuden piiriin kuuluvat tilanteet, joissa algoritmisia järjestelmiä hyödynnetään rikoksen tekemiseen ja mäin ollen ne vertautuvat työkaluun, ja akuuteista teknisistä ongelmista seuraavat yksittäiset ongelmat eli niin kutsutut bugit.

171 Sosiaalisen median konteksissa näiden tekijöiden vaikutuksesta ks. Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 225.

tallisia ilmiöitä ja luoda edellytyksiä uusien[172] haitallisten toimintojen syntymiselle[173]. Kun teknologioista kumpuavia haittoja tarkastelee, on ilmeistä, että haitta syntyy lopulta aina vuorovaikutussuhteessa, joissa merkityksellisiä elementtejä on useita. Teknologia on niistä vain yksi: teknologia on aina jonkun suunnittelemaa ja jonkun käyttöön ottamaa; sitä hyödyntää joku tavalla, joka mahdollisesti synnyttää haittoja jollekin, tai se vaikuttaa johonkuhun tavalla, jonka voi arvioida heikentävän tämän mahdollisuuksia elää hyvää elämää. Haittojen hallintaa samoin kuin haitan aiheuttajan yksiselitteistä tunnistamista vaikeuttaakin huomattavasti se, että nykymaailmassa ei ole helppoa tai edes mahdollista vetää rajaa virtuaalisen ja ei-virtuaalisen todellisuuden välille[174] tai erottaa teknologian vaikutusta muista vaikutuksista. Aina ei ole myöskään mahdollista tunnistaa, missä vaiheessa teknologian kehitystä tai käyttöä muodostuvat puitteet haitan syntymiselle. Jos ja kun haittoja ilmenee, niiden aiheuttajaa voi olla hyvin vaikea paikantaa. Lisäksi ihmisillä on hyvin erilaiset mahdollisuudet sopeutua muutoksiin, joita uudet teknologiat aiheuttavat yhteiskunnassa, jolloin haitat eivät välttämättä kosketa kaikkia, joihin teknologian vaikutukset kohdistuvat. Kuten jäljempänä tarkemmin esitän, läpitunkevat algoritmiset järjestelmät voivat kuitenkin paitsi uhata sekä yksilöiden että yhteisöjen oikeuksia[175], myös pahimmillaan heikentää sosiaalista oikeudenmukaisuutta[176] ja demokratiaa[177].

172 Tällaisista voi mainita esimerkkinä kohdennetun poliittisen vaikuttamisen, jonka algoritmiset teknologiat mahdollistavat ennennäkemättömällä tavalla, ks. esimerkiksi Zuiderveen Borgesius, F. J., Moeller, J., Kruikemeier, S., Fathaigh, R., Irion, K., Dobber, T., Bodó, B., & de Vreese, C. H. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, *14*(1), 82–89.

173 McQuillan, Dan. Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press, 2022;

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *AI Now Institute.*

174 Hildebrandt, M. (2015). Smart technologies and the end (s) of law: novel entanglements of law and technology. Edward Elgar Publishing.

175 Ks. esimerkiksi O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown;

Yeung, K. (2019). Why worry about decision-making by machine? Teoksessa *Algorithmic regulation*. Oxford University Press.

176 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

177 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3);

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (2020). Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse. *AlgorithmWatch*.

3.2 Haittojen typologiat

Haittoja on mahdollista tarkastella monesta näkökulmasta. Algoritmisiin teknologioihin liittyvien haittojen kohdalla on merkityksellistä tarkastella ainakin kolmea seikkaa: miten haitta ilmenee, miten teknologia tosiasiallisesti vaikuttaa haitan syntymisen taustalla ja kehen tai mihin haitta se kohdistuu.

Yhteiskunnallisten haittojen tutkimusperinteessä on tutkittu monipuolisesti kahta ensimmäistä: erilaisia haittojen ilmenemismuotoja ja haittojen kohdentumista. Sen sijaan haittojen taustalla olevien teknologioiden vaikutusten tutkimus on ollut verrattain vähäistä. Algoritmisten haittojen vaikutusmekaniikassa teknologioiden vaikutusten ajalliset ja maantieteelliset ulottuvuudet nousevat merkittäviksi tekijöiksi. Ne auttavat ymmärtämään, miksi ja miten osa algoritmisista haitoista ilmenee vasta kaukana algoritmisen teknologian välittömästä vaikutuspiiristä. Jotta pystyn huomioimaan tämän tarkastellessani algoritmisia haittoja, haen tukea tekoälyn ja teknologioiden tutkijoiden saavutuksista.

Kun algoritmisten järjestelmien erityispiirteet haittojen syntymekaniikassa ymmärretään ja tämä tieto tuodaan yhteiskunnallisten haittojen tutkimusperinteen tueksi, pystytään merkittävällä tavalla laajentamaan käsitystä seikoista, jotka vaikuttavat algoritmisten haittojen syntyyn ja leviämiseen. Seuraavissa alaluvuissa tarkastelen erilaisia tapoja luokitella haittoja ja niiden ominaisuuksia sekä toisaalta viitekehyksiä, jotka auttavat asemoimaan algoritmiset haitat yhteiskunnallisten haittojen tutkimuskentälle.

3.2.1 Haittojen ilmenemismuoto

Yhteiskunnallisten haittojen tutkijat ovat tutkineet laajasti niin sanottuja analogisia haittoja[178], joiden toteutumisessa algoritmiset teknologiat eivät ole määrittävässä asemassa[179]. Tällaisten haittojen luokittelemiseen on luotu erilaisia haittojen typologioita, joiden kautta haittojen ilmenemismuotoja voidaan analysoida. Typologioissa

[178] Näillä viitataan haittoihin, joiden syntymisessä algoritmiset teknologiat eivät ole välttämättä minkäänlaisessa roolissa. Vrt. algoritmiset haitat, jotka syntyvät nimenomaan algoritmisten teknologioiden käytön ja leviämisen seurauksena. Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 270–291.

[179] Ks. esimerkiksi Tombs, S. ja Hillyard, P. (2004). Towards a political economy of harm: States, corporations and the production of inequality. Teoksessa Hillyard, P., Pantazis, C., Tombs, S. & Gordon, D. (eds). *Beyond Criminology: Taking harm seriously*: 30–54. London: Pluto Press.

Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

Canning, V., & Tombs, S. (2021). From social harm to zemiology: A critical introduction. Routledge.

luokitellaan erityyppisiä haittoja, jotka heikentävät hyvinvointia ja siten mahdollisuuksia elää kukoistavaa elämää. Hyvinvoinnin ja inhimillisen kukoistuksen kannalta ajatellen merkitykselliset seikat ovat samat, vaikka niiden loukkaamisen tapa muuttuisi. Näin ollen on jokseenkin ilmeistä, että algoritmisten teknologioiden aiheuttamat haitat vastaavat tässä suhteessa pitkälti analogisia haittoja[180], ja samat typologiat soveltuvat siis melko hyvin myös algoritmisten haittojen arviointiin.

Esitin ensimmäisessä luvussa, että haittojen määrittely niiden ilmenemismuodon perusteella johtaa väistämättä myös siihen, että määritellään vähintään implisiittisesti, minkälainen on tavoiteltava tila tai *hyvä elämä,* jonka toteuttamista haitat estävät. Jotta hyvän elämän määrittely ei jäisi implisiittiseksi, tukeudun Nussbaumin[181] luokitteluun tekijöistä[182], jotka mahdollistavat hyvän elämän. Näin ollen voidaan argumentoida, että haitta on jotain, joka tosiasiallisesti rajoittaa, estää tai vaikeuttaa hyvän elämän edellytysten täyttymistä. Haittojen tutkimuksessa hyvän elämän edellytyksiä pyritään lisäämään tunnistamalla, analysoimalla ja lopulta myös rajoittamalla inhimillistä kukoistusta estäviä seikkoja sekä jäsentelemällä niiden syntyyn johtavia tekijöitä.

Yhteiskunnallisten haittojen tutkijat ovat esittäneet useita eri seikkoja painottavia haittojen typologioita. Näitä typologioita yhdistelemällä esitän, että haitoiksi voidaan lukea sellaiset tapahtumat tai laiminlyönnit, joiden seurauksena syntyy negatiivisia vaikutuksia 1) taloudelliseen turvallisuuteen, 2) terveyteen ja hyvinvointiin (fyysiseen tai psyykkiseen), 3) autonomiaan tai 4) mahdollisuuksiin ja asemaan yhteiskunnassa ja/tai suhteessa muihin (*relational harms*). Viimeisin käsittää esimerkiksi syrjinnän ja ekskluusion; väärintunnistamisen (*misrecognition),* joka viittaa yhtäältä siihen, että ihmisen viiteryhmän olemassaoloa ei huomioida tai siihen liittyviä piirteitä ei tunnisteta riittävällä tasolla, mikä voi johtaa esimerkiksi riittämättömiin tukitoimiin tai palveluihin; toisaalta kulttuurisen turvallisuuden heikentämiseen[183] ja sosiaalisen eriarvoisuuden voimistamiseen ja/tai ylläpitämiseen. Valitsemani jaot-

180 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

181 Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

182 Nussbaum nimeää kymmenen tekijää, joita hän kutsuu kyvykkyyksiksi: 1) Normaalin mittainen elämä, 2) fyysinen terveys, ravinto ja suoja, 3) ruumiillinen koskemattomuus, 4) aistit, mielikuvitus ja ajattelu, 5) tunteet, 6) kriittinen ajattelu, 7) yhteiselämä toisten ihmisten kanssa, 8) muut lajit, 9) leikki, 10) osallistuminen ja vaikuttaminen omaan ympäristöön, omistaminen. Ibid.

183 Alvesalo, A. (1999). Meeting the expectations of the local community on safety—what about white-collar crime? Konferenssiesitys *27th Annual Conference of the European Group for the Study of Deviance and Social Control*, Liettua, 2.–5.9.1999.

telu on yhdistelmä Pembertonin[184] sekä Hillyardin ja Tombsin[185] haittojen typologioita, ja siinä tiivistyvät Nussbaumin esittämät hyvän elämän edellytykset[186]. Hyödynnän jaottelua, kun pyrin selvittämään, *minkälaisia* haittoja tekoäly- ja algoritmiset järjestelmät voivat saada aikaan.

Haittojen typologiat lähtevät siitä premissistä, että haittoja on mahdollista luokitella tiettyihin ryhmiin sen mukaan, mihin inhimillisiin intresseihin ne kohdistuvat. Tällaisessa luokittelussa kiinnostus kohdistuu siis erityisesti siihen, minkälainen haitta on. Sen sijaan haittojen syntytapa ei siis ole määrittävä tekijä: olennaista on tunnistaa, miten haitat heikentävät mahdollisuuksia elää kukoistavaa elämää. Algoritmisten haittojen erityispiirteiden tunnistamiseksi täytyy kuitenkin tarkastella myös, miten algoritmiset haitat syntyvät.

3.2.2 Haittojen syntymekaniikka

Haittojen syntymekaniikan osalta algoritmiset haitat eroavat huomattavalla tavalla analogisista haitoista. Analogisten haittojen synnyn voi nähdä kiinnittyvän melko kiinteästi ihmisten tekemiin päätöksiin, ratkaisuihin ja toimintoihin. Vaikka käytössä voi olla erilaisia teknologisia apuvälineitä, jotka vaikuttavat haittojen syntymiseen, analogiset haitat syntyvät pääsääntöisesti jokseenkin tiiviissä yhteydessä inhimilliseen toimijaan tai toimijoihin. Sen sijaan algoritmisten haittojen kohdalla etäisyys yksittäisen inhimillisen teon ja syntyneen haitan välillä voi olla huomattava.

Algoritmisten haittojen syntymiseen liittyvää problematiikkaa on tutkittu jokseenkin vähän. Haittoja syntyy erilaisten prosessien seurauksena ja se, kuinka suoraviivaisesti haitat ilmenevät, vaihtelee. Haitat voivat syntyä paitsi välittömästi jonkin toiminnan tai päätöksen seurauksena myös epäsuorasti tai vähittäin erilaisten vähäisten muutosten kertautuessa ja kerrostuessa. Tämä tuo haittojen arviointiin olennaisen muuttujan: etäisyyden. Algoritmisia haittoja onkin arvioitava myös suhteessa aikaan ja tilaan. Haittojen etäisyys teknologiasta voi kasvaa erityisesti tilanteissa, joissa algoritmisten teknologioiden näkymätön tai huomaamaton toiminta vähittäin muovaa yhtäältä käyttäjiään ja toisaalta ympäristöään ja siinä toimivia tahoja.

184 Pemberton, S. A. (2015). Harmful societies: Understanding social harm. Policy Press.

185 Tombs, S. ja Hillyard, P. (2004). Towards a political economy of harm: States, corporations and the production of inequality. Teoksessa Hillyard, P., Pantazis, C., Tombs, S. & Gordon, D. (eds). *Beyond Criminology: Taking harm seriously*: 30–54. London: Pluto Press.

186 Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

Wood[187] on analysoinut laaja-alaisesti ihmisten, teknologioiden ja haittojen välisiä suhteita. Hänen luomansa viitekehys toimii lähtökohtana, kun tarkastelen algoritmisen teknologian ja syntyneen haitan välistä suhdetta ja etäisyyttä. Kun viitekehys tuodaan yhteiskunnallisten haittojen teorian tueksi, se mahdollistaa kattavasti myös epäsuorien, kaukana teknologiasta ilmenevien haittojen analysoimisen. Lisäksi kehys antaa mahdollisuuden tarkastella teknologisen järjestelmän odottamattomia tai pahantahtoisia käyttötapoja ja niistä kumpuavia haittoja.

Wood hahmottelee teknologioihin liittyviä haittoja jakamalla ne yhtäältä sen mukaan, onko teknologian käyttötapa sen tarkoituksen mukainen eli suunniteltu vai siitä poikkeava eli odottamaton, ja toisaalta sen mukaan, kiinnittyykö haitta suoraan teknologian (suunniteltuun tai odottamattomaan) toimintaan tai käyttöön vai onko se teknologian *vaikutuksesta* johtuva. Tästä muodostuu nelikenttä: haitta voi syntyä teknologian suunnitellun tai odottamattoman käytön/toiminnan välineellisenä seurauksena tai generatiivisena vaikutuksena. *Välineelliset haitat* viittaavat siihen, että teknologiaa aktiivisesti käytetään tavalla, joka aiheuttaa haitallisen vaikutuksen: teknologia on siis väline, jolla haitta tuotetaan. *Generatiiviset haitat* taas viittaavat teknologioiden potentiaaliin muuttaa joko niiden käyttäjää tai käyttöympäristöä tavoilla, jotka muokkaavat todellisuutta haitalliseen suuntaan.

Taulukko 2. Haittojen nelikenttä Woodin mukaan.[188]

	väline (instrumental)	**vaikutus (generative)**
Suunniteltu käyttö/toiminta (utility)	Suunnitellun käytön/ toiminnan välineelliset haitat	Suunnitellun käytön/ toiminnan generatiiviset haitat
Odottamaton käyttö/toiminta (technicity)	Odottamattoman käytön/ toiminnan välineelliset haitat	Odottamattoman käytön/ toiminnan generatiiviset haitat

Viitekehys on jokseenkin tekninen ja sinänsä vaikeasti lähestyttävä, ja sitä on toistaiseksi hyödynnetty vasta vähän yhteiskunnallisten haittojen tutkimuksessa. Sen avulla teknologioista johtuvia haittoja pystytään kuitenkin tunnistamaan ja tarkastelemaan laajemmin kuin perinteisesti on yhteiskunnallisten haittojen tutkimuksessa totuttu tekemään. Viitekehys mahdollistaa teknologian vaikutuksista juontuvien haitallisten seurausten palauttamisen takaisin teknologiaan, mikä lisää mahdollisuuksia tunnistaa algoritmisia haittoja, jotka ilmenevät ajallisesti tai maantieteellisesti etäällä teknologian välittömästä vaikutuspiiristä.

187 Wood, M. A. (2021). Rethinking how technologies harm. *The British Journal of Criminology*, *61*(3), 627–647.

188 Wood, M. A. (2021). Rethinking how technologies harm. *The British Journal of Criminology*, *61*(3), 627–647.

Woodin viitekehyksessä voidaan nähdä yhtymäkohtia myös yhteiskunnallisten haittojen tutkimuksen parissa tutkittuihin heijastusvaikutuksiin (*ripple effect*)[189]. Heijastusvaikutukset ovat haittoja, jotka saavat alkunsa haitallisesta tapahtumasta ja leviävät siitä sekä maantieteellisesti että ajallisesti yhä laajemmalle alueelle. Wood puhuu toisen asteen haitoista (second order harm) vastaavassa merkityksessä: ne ovat ensisijaisesta haitasta seuraavia uusia haittoja[190]. Generatiiviset haitat sen sijaan ovat haittoja, joita teknologian käyttämisestä seuraa, kun se vaikuttaa ympäristöönsä tai käyttäjäänsä.

Huomioidakseni laajasti algoritmisiin teknologioihin liittyvän haittapotentiaalin, jaottelen haitat välittömiin ja seurannaishaittoihin. Välittömät haitat ilmenevät tekoälyteknologian välittömässä vaikutuspiirissä ja palautuvat suoraan teknologian toimintaan. Sen sijaan seurannaishaitat ilmenevät teknologian toiminnan seurauksena, mutta etääntyvät teknologiasta, eli teknologian vaikutus haitan syntyyn voi olla tilassa ja/tai ajassa epäsuoraa. Jaossa on paljon samaa Woodin nelikentän kanssa. Välittömät haitat rinnastuvat jokseenkin suoraviivaisesti Woodin määritelmän mukaisiin välineellisiin haittoihin. Seurannaishaitat pitävät kuitenkin sisällään Woodin argumentaatiossa esiin nousevien generatiivisten haittojen lisäksi myös yhteiskunnallisten haittojen tutkimusperinteessä määritellyt heijastusvaikutukset, joita Wood nimittää toisen asteen haitoiksi.

Woodin jaottelussa merkityksellistä on teknologian rooli haitan aiheuttamisessa, jolloin arvioinnin kohteena on, millä tavalla haitta kytkeytyy teknologiaan. Omassa jaottelussani välittömiin ja seurannaishaittoihin kiinnekohtana on teknologian ja ilmenevän haitan välinen etäisyys niin ajassa kuin tilassa: kuinka lähellä haitan synnyn mahdollistavaa seikkaa haitta ilmenee. Tämä mahdollistaa viitekehyksen soveltamisen myös tilanteissa, jossa haitallinen kehityskulku käynnistyy jonkun muun seikan kuin teknologian vaikutuksesta. Käytännössä seurannaishaitat lähestyvät heijastusvaikutuksia, mutta niiden katalysaattorina voi olla välittömiä haittoja tuottavan tapahtuman lisäksi myös positiivisina tai neutraaleina koetut muutokset. Esimerkiksi sinänsä toivottu ja tarpeellinen informaatiotulvan rajoittaminen erilaisilla algoritmisilla suodattimilla paitsi parantaa ihmisten mahdollisuuksia löytää haluamaansa tie-

189 Tombs, S. (2019). Grenfell: The unfolding dimensions of social harm. *Justice, Power and Resistance*, *3*(1), 61–88.

190 Voidaan esimerkiksi kuvitella tilanne, jossa algoritminen järjestelmä epää hakijalta rahallisen tuen, johon tämä olisi oikeutettu. Tuen puuttuminen itsessään aiheuttaa taloudellista haittaa, joka voi heikentää hakijan mahdollisuuksia toimia yhteiskunnassa. Hakijan heikko taloudellinen tilanne voi heijastua muihinkin. Voi olla, että hakijalla ei ole varaa kustantaa lapsensa harrastusta, mikä voi heikentää myös lapsen fyysistä ja/tai psyykkistä hyvinvointia. Seuraukset voivat olla pitkäkestoisia. Yksittäinen haitallinen tapahtuma siis voi heijastaa vaikutuksensa paitsi ajassa, myös tilassa laajemmalle alueelle.

toa, myös heikentää ihmisten tiedonsaantimahdollisuuksia ja siten vinouttaa ihmisten tietämystä[191]. Tällä on vaikutuksia ihmisen kykyyn tehdä perusteltuja ratkaisuja, minkä seurauksena ihmisen autonomia kärsii.

Algoritmisia seurannaishaittoja ovat myös esimerkiksi välittömien algoritmisten haittojen aiheuttamat haitalliset vaikutukset, epäsuorien – tai Woodin termistöä hyödyntäen generatiivisten – algoritmisten haittojen aiheuttamat haitalliset vaikutukset, teknologioiden sinänsä toivotusta toiminnasta seuraavat haitalliset vaikutukset ja erilaisten teknologioiden yhteisvaikutuksista syntyvät haitalliset vaikutukset. Erilaisiin algoritmisiin seurannaishaittoihin palataan jäljempänä tarkemmin.

3.2.3 Haitan kohde

Algoritmisen haitan *ilmenemismuodon* sekä ilmenneen *haitan syntymekaniikan* arvioimisen lisäksi on luonnollisesti olennaista huomioida myös, keneen tai mihin haitta kohdistuu. *Haitan kohde* onkin kolmas algoritmisten haittojen luokittelemisen ja analysoimisen kannalta olennainen tekijä.

Yhteiskunnallisten haittojen tutkimuksen parissa on perinteisesti arvioitu haittoja, jotka kohdistuvat yksilöihin ja/tai ryhmiin. Tässäkin tutkimuksessa tällaiset haitat ovat merkittävässä roolissa. Sen sijaan *yhteiskuntaa* ei ole aiemmin juurikaan nähty haittojen kohteena; päinvastoin aiemmassa tutkimuksessa yhteiskunta on nähty jokseenkin pelkästään haittoja synnyttävänä rakenteena. Ymmärrän lähtökohdan, mutta pidän sitä etenkin transformatiivisia teknologioita koskevissa kysymyksissä jossain määrin puutteellisena.

Jos katsomme Giddensin rakenteistumisteorian[192] mukaisesti, että yhteiskunnan rakenteen ja inhimillisen toimijan, yksilön, vuorovaikutussuhde on duaalinen eli yhteiskunta ja toimija ovat toistensa määrittäviä tekijöitä, yhteiskunnan sivuuttaminen haittojen kohteena vaikuttaa jättävän suuren osan potentiaalisia haittoja huomiotta. Giddensin mukaan yksilön kaikki toiminta uusintaa, heikentää tai muuttaa sosiaalisia käytänteitä, jotka vaikuttavat sosiaalisiin systeemeihin ja niiden kautta yhteiskunnan näkymättömiin sääntöihin, rakenteeseen. Yhteiskunnan rakennetta on siis Giddensin teorian mukaan mahdollista muuttaa – ja tarkoituksellisen ja tavoitteellisen muuttamisen lisäksi rakenne väistämättä *muuttuu* myös yksilöiden toimien tahattomana seurauksena. Tahattoman muutoksen suunta voi olla täysin odottamaton. Samanaikaisesti yhteiskunnan rakenne kuitenkin määrittää yksilön toiminnalle reunaehdot,

191 Puhutaan kuplista (*filter bubble*) ja kaikukammioista (*echo chamber*). Ks. Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298–320.

192 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

jolloin yhteiskunnan rakenteen muutokset väistämättä muokkaavat yksilöiden toiminnan mahdollisuuksia ja siis myös mahdollisuuksia hyvän elämän tavoittelussa ja saavuttamisessa. Jos hyväksymme tämän lähtökohdan, voimme ymmärtää, miksi yhteiskuntaa on merkityksellistä tutkia myös haittojen kohteena. Tällöinkin on kuitenkin selvää, että lopullinen, konkreettinen haitta näkyy inhimillisen toimijan elämässä. Yhteiskunnan muutos haitallisempaan ainoastaan luo perustan haitan syntymiselle. Yhteiskuntaan kohdistuvat haitat siis lisäävät todennäköisyyttä yksilöihin ja yhteisöihin kohdistuvien haittojen syntymiselle, kun yhteiskunnan toiminnan mahdollisuudet heikkenevät.

Seuraavassa alaluvussa keskiössä ovat haitat, joita algoritmisista teknologioista on seurannut. Käsitteellistän niitä hyödyntämällä yllä kuvaamiani haittojen luokittelun tapoja. Tekoälykentän laajuuden takia ei ole perusteltua pyrkiä listaamaan kaikkia potentiaalisia haittoja, joten nostan esimerkinomaisesti esille algoritmisiin järjestelmiin tai niiden käyttöön liittyviä seikkoja, jotka ovat merkittäviä, kun arvioidaan algoritmisten teknologioiden merkitystä yhteiskunnallisten haittojen kontekstissa.

3.3 Yksilöön kohdistuvat haitat

3.3.1 Välittömät haitat

Tässä alaluvussa käsittelen välittömästi algoritmisiin teknologioihin kiinnittyviä, yksilöön kohdistuvia haittoja, joita etenkin tekoäly- ja algoritmitutkimuksen perinteissä on tunnistettu. Tarkastelen yhtäältä, millaisia välittömiä haittoja algoritmisten teknologioiden käytöstä seuraa, ja toisaalta, minkälaiset seikat haittojen syntyyn vaikuttavat. Koska välittömät haitat nimensä mukaisesti seuraavat välittömästi teknologian toiminnasta tai käytöstä, haittojen syntyä edesauttavien seikkojen voi olettaa suurelta osin kiinnittyvän teknologioiden toiminnan tai käytön tapoihin. Nämä teemat ovat luvussa suuressa roolissa. Tarkastelun perusteella voidaan arvioida, miten tekoälyteknologioiden muuttavat haittojen syntymisen ja leviämisen mahdollisuuksia ja mekaniikkaa aiemmasta.

3.3.1.1 Haittojen ilmenemismuodot

Yksilöihin kohdistuvia, tekoälyjärjestelmien luomia välittömiä haittoja on tunnistettu jo paljon. Tällaiset haitat kiinnittyvät pääsääntöisesti teknologioiden ”virheellisiin” päätöksiin. Virheellisyys ei viittaa niinkään siihen, että järjestelmä itsessään toimisi virheellisesti eli ohjelmointinsa vastaisesti vaan siihen, että sen toiminnan lopputulos ei vastaa suunniteltua tai tavoiteltua.

Epätarkoituksenmukaisesta toiminnasta johtuu monenlaisia haittoja, sillä algoritmisia teknologioita hyödynnetään mitä moninaisemmissa yhteyksissä. Fyysisten

tekoälyjärjestelmien, kuten autonomisten autojen tai erilaisten robottien, virheelliset, odottamattomat tai epätarkoituksenmukaiset päätökset voivat johtaa fyysisiin onnettomuuksiin, kun taas virtuaalisessa ympäristössä operoivien tekoälyjärjestelmien epätoivottu toiminta voi heikentää monin eri tavoin ihmisten elämisen ja toiminnan mahdollisuuksia. Taloudelliseen turvallisuuteen[193] vaikuttavat esimerkiksi monet hallinnon ja rahoitusalan tekoälyjärjestelmät ja terveyteen[194] terveydenhuollon diagnostiikka-algoritmit. Algoritmiset järjestelmät vaikuttavat monia reittejä myös ihmisten autonomiaan ja toimintamahdollisuuksiin yhteiskunnassa: algoritmisia järjestelmiä hyödynnetään läpi yhteiskunnan analysoinnin, profiloinnin ja ennusteiden tekemisen tukena, ja yhä useammin päätöksentekojärjestelmät toimivat myös vailla ihmisen välitöntä kontrollia. Algoritmit määrittävät näin ollen yhä laajemmin ihmisten mahdollisuuksia[195].

Näyttää siltä, että järjestelmän toiminnan epätarkoituksenmukaisuus voi seurata joko järjestelmän suunnittelun tai tuotannon virheistä tai siitä, että järjestelmä ei sovi siihen käyttötarkoitukseen tai -ympäristöön, missä sitä hyödynnetään. Jälkimmäinenkin vaihtoehto kuitenkin palautuu osittain takaisin kysymyksiin teknologian toteutuksesta, sillä myös epäsopivat käyttötavat määräytyvät niiden valintojen perusteella, joita teknologian tuotantovaiheessa tehdään.

3.3.1.2 Dataan liittyvistä kysymyksistä

Koneoppivien järjestelmien kohdalla yksi suurimmista haittoihin kytkeytyvistä teemoista on data, jota järjestelmien opetukseen käytetään. Dataan liittyvää tutkimusta on tehty laajasti niin oikeustieteen[196], yhteiskuntatieteiden[197] kuin teknisten tieteenalojen[198] parissa. Datan avulla automaattiset päätöksentekojärjestelmät voidaan koneoppimismenetelmin opettaa ennustamaan todennäköisimmin oikeaa ratkaisua.

193 Algoritmeja on hyödynnetty esimerkiksi Yhdysvaltojen Michiganissa sen arvioimiseen, täyttääkö henkilö kriteerit työttömyyskorvauksen saamiseksi. Ks. Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

194 Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347.

195 Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

196 Ks. esimerkiksi Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

197 Ks. esimerkiksi Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

198 Ks. esimerkiksi O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Käytännössä järjestelmän tuottama ratkaisu perustuu usein historialliseen dataan, jonka perusteella järjestelmä on opetettu. Luonnollisesti tällainen järjestelmä voi toimia halutulla tavalla ainoastaan silloin, kun sitä käytetään tilanteissa, joissa järjestelmän tekemien ratkaisujen perusteena on opetusdataan riittävän hyvin verrannollinen materiaali eli opetusdata vastaa riittävällä tasolla uutta dataa, jota järjestelmälle syötetään. Järjestelmän tekemien ratkaisujen laatu kiinnittyy siis vahvasti, joskaan ei pelkästään, kysymyksiin järjestelmän sopivuudesta ympäristöönsä ja valintoihin, joita opetusdatan keräämisen, rajaamisen ja käsittelyn aikana on tehty.

Monien laajasti tunnistettujen välittömien algoritmisten haittojen taustalla voidaan nähdä opetusdataan liittyvää problematiikkaa[199]. Algoritmisten järjestelmien opetusdatasta seuraavat haitat voivat perustua eri tekijöihin. Esimerkiksi Mittelstadt ja muut[200] argumentoivat, että algoritmisten järjestelmien toimintaa ohjaavien tilastollisten tai koneoppivien menetelmien avulla ei ole lähtökohtaisesti mahdollista saada varmaa tietoa, sillä tarkoituksenmukaisenkin datan tulkitsemisessa on väistämättä mukana tietty määrä epävarmuutta (inconclusive evidence). Tämä epävarmuus voi järjestelmän käytössä realisoitua virheellisinä tai epätoivottuina ratkaisuina. Toisaalta on myös riski, että ei ole riittävää tietoa tai ymmärrystä siitä, minkälaisella datapohjalla järjestelmä tosiasiallisesti toimii, jolloin ei voida olla varmoja, mitä tai mistä järjestelmän tuottama lopputulos kertoo (inscrutable evidence). Lopputulos, jonka uskotaan kertovan jotain, mitä se ei todellisuudessa voi kertoa, voi johtaa haitallisiin tulkintoihin tai ratkaisuihin. Lisäksi on selvää, että datapohjaisen järjestelmän tuottama lopputulos on väistämättä korkeintaan niin hyvä kuin datapohja, jonka perusteella järjestelmä operoi. Opetusdatassa tiivistyvät herkästi yhteiskunnan rakenteisiin kiinnittyneet ja historian saatossa toistuneet vääristymät, ennakkoluulot ja epäoikeudenmukaisuudet (misguided evidence). Kun nämä siirtyvät algoritmisen järjestelmän toimintaan, puhutaan vinoumista (bias).

Vinouma viittaa siis vääristymään. Aihepiirin tutkijoista monet katsovat, että kaikki ennustemallit ja koneoppivat algoritmiset järjestelmät ovat enemmän tai vähemmän vinoutuneita[201]. Vinoutunut algoritminen järjestelmä tuottaa vääristyneitä, mahdollisesti epäoikeudenmukaisia (*algorithmic unfairness*) ratkaisuja. Jos tällaisia järjestelmiä käytetään päätöksentekoon tai päätöksenteon tukena, ne voivat esimerkiksi perustaa päättelynsä yksilön sellaisiin ominaisuuksiin, joilla ei ole tai joilla ei saisi olla päätöksen kannalta merkitystä. Tämän seurauksena päätöksentekoproses-

199 Ks. esimerkiksi O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

200 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

201 Ks. esimerkiksi Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs,* 49(2), 209–231.

sissa päädytään herkästi virheellisiin ja mahdollisesti syrjiviin lopputuloksiin[202]. Vastaavaan logiikkaan tukeutuen esimerkiksi McQuillan argumentoi, että tekoäly on väistämättä fasistinen työkalu, joka ylläpitää ja ruokkii yhteiskunnassa esiintyvää rakenteellista eriarvoisuutta[203]. Onkin selvää, että data, jonka perusteella koneoppivat tekoälyjärjestelmät koulutetaan, ei voi syntyä irrallaan yhteiskunnallisesta todellisuudesta. Opetusdatassa toistuva logiikka, joka mitä todennäköisimmin heijastelee yhteiskunnassa vallitsevia rakenteita, siirtyy väistämättä sellaisenaan tekoälyjärjestelmien päättelyyn, mikäli eriarvoisuutta ylläpitäviä rakenteita ja niiden ilmenemistä datassa ei tunnisteta ja niihin puututa.

Sen lisäksi, että yhteiskunnallinen todellisuus ja sen rakenteellinen eriarvoisuus väistämättä näkyvät opetusdatassa, ne näkyvät myös datan *keräämisessä.* Epätasa-arvo siis vaikuttaa dataan ja sitä kautta valuu tekoälyjärjestelmien toimintaan monia eri reittejä. Epätasa-arvo ja rakenteellinen eriarvoisuus kiinnittyvät poliittisiin valintoihin, valtaan ja vallankäyttöön eli teemoihin, joilla on valtava vaikutus myös datan keräämisen ja käyttämisen kontekstissa. Se, millaista dataa päätyy tekoälyjärjestelmien opetuskäyttöön, riippuu yhtäältä poliittisten päätösten ohjaamana luoduista lainsäädännöllisistä rajoista, toisaalta käytänteistä, jotka ovat syntyneet tietyssä yhteiskunnallisessa todellisuudessa tehtyjen, yhteiskunnan reunaehtoihin sopivien – tai niitä haastavien – ja siis väistämättä poliittisten ratkaisujen seurauksena[204]. Datan keräämistä ja hyödyntämistä siis sääntelee aina viimesijassa valta[205]: kuka voi kerätä dataa, mitä dataa voidaan kerätä, kenen dataa kerätään ja mihin sitä käytetään. Data ei siis ole neutraalia[206], sillä sen olemassaolo kiinnittyy kysymyksiin siitä, kenelle valta datan keräämiseen ja käyttämiseen kuuluu. Monet dataan liittyvistä konkreettisista kysymyksistä palautuvatkin kysymyksiin siitä, kenellä on oikeus kerätä ja kenen dataa, mistä data kertoo ja mihin sitä hyödynnetään. Tehtyihin valintoihin liittyy väistämättä myös riskejä datan epäoikeudenmukaisesta käyttämisestä, datavalintojen epätarkoituksenmukaisuudesta sekä opetusdatan puutteellisuudesta, vinoumista ja myös tarkoituksellisesta vääristelystä[207]. On osoitettu, että valinnat, joita datan ke-

202 Ojanen, A., Sahlgren, O., Vaiste, J., Björk, A., Mikkonen, J., Kimppa, K., Laitinen, A. & Oljakka, N. (2022). Algoritminen syrjintä ja yhdenvertaisuuden edistäminen: Arviointikehikko syrjimättömälle tekoälylle. Valtioneuvoston kanslia.

203 McQuillan, D. (2022). Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press.

204 Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

205 McQuillan, D. (2022). Resisting AI: an anti-fascist approach to artificial intelligence. Policy Press.

206 Ibid. 94.

207 Riskit on tiedostettu kauan myös valtioiden tasolla. Ks. esimerkiksi Yhdysvalloista Executive Office of the President. (2014). *Big Data: Seizing Opportunities, Preserving Values.*

räämisen, muokkaamisen ja käytön aikana tehdään, voivat johtaa tekoälyjärjestelmän virheelliseen tai epätarkoituksenmukaiseen toimintaan, kuten vinoutuneisiin tai syrjiviin päätöksiin[208]. Selvää on, että virheellinen toiminta kasvattaa riskiä myös haittojen syntymiselle.

3.3.1.3 Algoritmisten teknologioiden vaikutukset päätöksentekoprosesseihin

Dataan liittyvien kysymysten merkittävyydestä huolimatta algoritmisiin järjestelmiin liittyvät, yksilöön kohdistuvat välittömät haitat eivät pelkisty niihin. Erityisesti automaattisen päätöksenteon kontekstissa on selvää, että tuotantovaiheessa tehdyt data- ja muut valinnat vaikuttavat siihen, kuinka asianmukaisia päätöksiä järjestelmä käytössä ollessaan voi tehdä. Asianmukaisestikin toimivat algoritmiset teknologiat kuitenkin muuttavat päätösten tekemisen mekaniikkaa tavoilla, jotka voivat johtaa haittojen syntymiseen. Etenkin hallinnollisessa päätöksenteossa, mutta myös monissa muissa tilanteissa, joissa tehdään ihmisten elämään vaikuttavia päätöksiä, niiden oikeutus perustuu paitsi yhteiskunnassa tunnettuihin ja hyväksyttyihin tavoitteisiin, myös päätöksen tekemisen tapoihin: siihen, että päätöksen taustalla on asiantuntijoiden harkinta, ja harkinnassa noudatetut perusteet tuodaan päätöksen kohteelle tai jopa julkisesti nähtäville[209]. Mikäli näin ei tapahdu, päätösten oikeuttaminen voi olla vaikeaa tai mahdotonta.

Tunnettuihin ja hyväksyttyihin tavoitteisiin perustuvien, yksilöllisten ratkaisujen sijaan algoritmiset järjestelmät soveltavat systemaattisesti samaa logiikkaa kaikissa tapauksissa, ja lopulliseen ratkaisuun johtava koneellinen päättely voi paeta inhimillistä ymmärrystä. Tämä lisää päätösten yhdenmukaisuutta, mutta samalla kasvattaa riskiä sille, että päätöksissä sovelletaan systemaattisesti virheellistä logiikkaa tai päätöksenteon joustavuus menetetään. Monien algoritmisten järjestelmien kohdalla toimintalogiikan läpinäkymättömyys vaikeuttaa virheiden tunnistamista ja niihin puuttumista[210].

Toimintamekaniikan muuttuminen ei ole tahatonta. Algoritmisilla järjestelmillä on ominaisuuksia, joiden avulla toimintaa voi tehostaa: valtava toimintakapasiteetti ja nopeus, kyky soveltaa koodinsa mukaisia perusteita luotettavasti jokaisessa ta-

208 Muiden muassa O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

209 König, P. D., & Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. *Technology in Society*, *67*, 101688.

210 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

pauksessa, käytön edullisuus. Ne mahdollistavat ihmisten työtaakan keventämisen, ja niiden avulla voidaan suoriutua nopeammin esimerkiksi päätöksenteosta, diagnostiikasta, asiakaspalvelusta ja monista muista toiminnoista. Kun ihmistyö tai osa siitä korvataan algoritmisella järjestelmällä, toiminta vaatii vähemmän resursseja, yhdenmukaistuu ja tehostuu.

Automaatioprosesseihin voi kuitenkin liittyä haittoja monilla tasoilla. Välittömien haittojen kannalta katsoen ensisijaisena riskinä on, että automaatio heikentää toiminnan, esimerkiksi päätöksenteon tai sen perusteluiden laatua, ja virheelliset tai mahdollisesti syrjivät ratkaisut lisääntyvät[211]. Koneellisessa päättelyssä inhimillinen ajattelu ja harkinta korvataan sääntöjen mukaisella päättelyllä tai todennäköisyyksien arvioimisella[212]: merkitystä annetaan vain sellaisille seikoille, jotka järjestelmän ohjelmointi- tai opetusvaiheessa on määritelty merkityksellisiksi. Tämä tarkoittaa, että ratkaisujen joustavuus vähenee. Joustavuuden vähentyminen voi johtaa käytänteiden jähmettymiseen. Tämä voi heikentää etenkin niiden henkilöiden oikeuksia ja mahdollisuuksia, joiden asia inhimillisessä harkinnassa käsiteltäisiin rajatapauksena.

Toisaalta, jos ja kun automaattisten järjestelmien käyttöönotto johtaa inhimillisen työvoiman vähentämiseen, se voi myös lisätä työttömyyttä ja aiheuttaa siihen liittyviä haittoja[213], esimerkiksi heikentää yksilöiden taloudellista turvallisuutta tai yhteiskunnallista asemaa. Algoritmisten järjestelmien käytöstä seuraavat muutokset voivat olla hyvin kauaskantoisia. Kysymys ei ole pelkästään välittömästi järjestelmän toiminnan seurauksena ilmenevistä välittömistä haitoista, jotka nekin voivat olla vakavia, vaan myös pidemmälle leviävistä seurauksista, jotka voivat pahimmillaan murentaa yksilön toimintakyvyn, heikentää luottamusta yhteiskunnallisiin järjestelmiin, kiristää yhteiskunnassa vallitsevia asenteita ja arvoja ja muuttaa yhteiskunnan toiminnan edellytyksiä. Tällaiset haitat lukeutuvat seurannaishaittoihin, joita käsitellään tämän luvun myöhemmissä osissa.

3.3.1.4 Algoritmisten haittojen syntymisen mekaniikasta

Algoritmiset, yksilöihin kohdistuvat välittömät haitat vaikuttaisivat ilmenevän vastaavissa muodoissa kuin analogisetkin haitat, mutta ne syntyvät ja leviävät eri tavoin

211 Chiusi, F., Fischer, S., Kayser-Bril, N., & Spielkamp, N. (2020). Automating society report 2020. AlgorithmWatch. 160–172.

212 Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197–211.

213 Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: how technology changes labor demand. *Journal of Economic Perspectives*, *33*(2), 3–30.

kuin analogiset vastineensa[214]. Algoritmisten ja analogisten haittojen synnyn ja leviämisen erot näyttäisivät perustuvan muutamaan olennaiseen seikkaan. Ensimmäinen on ihmistoimijan ja haitan välinen etäisyys. Algoritmisten haittojen ilmenemisen kannalta suuri osa merkittävistä inhimillisistä päätöksistä tehdään jo algoritmisen teknologian suunnitteluvaiheessa, ja käytössä ollessaan järjestelmä voi toimia jokseenkin itsenäisesti ja – teknisistä ratkaisuista riippuen – mukautua ympäristöönsä. Esimerkiksi opetusdatasta juontuvat haitat alustetaan jo opetusdatavalintojen ja datan käsittelyn yhteydessä eli paljon ennen kuin järjestelmä todellisuudessa tekee yhtäkään päätöstä. Vaikka mahdollinen algoritminen haitta ilmenisikin välittömästi algoritmisen järjestelmän toiminnan seurauksena, haittaan johtanut inhimillinen toiminta on tapahtunut huomattavasti aiemmin. Algoritmisen teknologian suunnittelijat ja tuottajat arvioivat siis teknologian toiminnan vaikutuksia ennalta, ajallisesti ja käytännöllisesti irrallaan teknologian tosiasiallisesta käytöstä. Tämä voi lisätä riskejä virheille, epätarkoituksenmukaisuudelle ja niistä johtuville haitoille.

Toinen erottava tekijä kiinnittyy algoritmisten järjestelmien toiminnan nopeuteen ja eri järjestelmien vuorovaikutukseen[215]. Algoritmisiin järjestelmiin voi kertyä tietoa monista eri lähteistä, ja eri järjestelmät voivat linkittyä tiiviisti yhteen. Mikäli näin tapahtuu, algoritmisten järjestelmien tuottamat ennusteet ja niihin pohjautuva mahdollinen päätöksenteko voivat käydä yksilölle vaikeasti ymmärrettäväksi, etenkin jos tietoa eri tietokantoihin kerätystä datasta tai järjestelmien välisistä yhteyksistä ei syystä tai toisesta ole[216]. Monimutkaiset ja toisiinsa liittyvät teknologiat luonnollisesti vähentävät yksilön mahdollisuuksia ymmärtää ja vaikuttaa asioihin, joita tällaiset teknologiat määrittävät, mikä näkyy erityisesti autonomiahaittoina. Lisäksi, kun algoritmiset teknologiat pystyvät toimimaan ja tekemään päätöksiä huomattavasti inhimillisiä toimijoita nopeammin, myös mahdollisten haittojen ilmenemistahti voi kiihtyä. Eri teknologioiden ketjuuntuvat ja yhteisvaikutukset lisäävät haittojen leviämisen potentiaalia: yhden järjestelmän virheellinen toiminta tai yksittäinen virheellinen kirjaus päätöksen kohteesta voi aiheuttaa ketjureaktion, jossa yksittäinen virhe moninkertaistuu. Tämä voi johtaa siihen, että virheitä ilmenee yhä uusissa yhteyksissä, jolloin myös haittapotentiaali kasvaa ja haittoja voi syntyä entistä useammissa muodoissa.

Haittapotentiaalia kasvattaa kolmas olennaisesti algoritmisiin teknologioihin liittyvä seikka. Algoritmisissa järjestelmissä monia analogisille järjestelmille tyypilli-

214 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

215 Ibid. 192.

216 Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

siä, turvallisuutta lisääviä mekanismeja ei ole. Siinä missä analogiset eli pääasiallisesti inhimillisten toimijoiden varassa toimivat järjestelmät ovat usein hajautettuja, ja monien inhimillisten toimijoiden päätökset vaikuttavat niiden toimintaan ja lopputulemiin, algoritmisille järjestelmille on tyypillistä keskitetty toiminta, systemaattisuus ja inhimillisten osatekijöiden vähäisyys toiminnan käynnistyttyä[217]. Tämä heikentää mahdollisuuksia tunnistaa ongelmia, puuttua niihin ja muokata käytänteitä paremmiksi.

Neljäs erottava tekijä on haittoja aiheuttavan toiminnan olemassaolon ja/tai järjestelmän toimintalogiikan hämärtyminen vaikeasti havaittavaksi[218]. Algoritmisille teknologioille usein tyypilliset ominaisuudet, kuten läpitunkevuus, huomaamattomuus ja toimintalogiikan läpinäkymättömyys, tekevät joistakin algoritmisten teknologioiden aiheuttamien haittojen havaitsemisesta ja niiden syiden selvittämisestä vaikeaa. Jos tekoälyteknologioita käytetään päätöksentekoon tai päätöksenteon tukena, eikä yksilö välttämättä tiedä, mistä kaikkialta häntä koskevaa tietoa kerätään tai miten sitä yhdistellään, seurauksena voi olla päätöksenteon arvaamattomuuden lisääntyminen. Se heikentää mahdollisuuksia varautua tulevaan, mikä voi itsessään aiheuttaa autonomiahaittoja. Kun algoritmisaatio on nopeaa, myöskään järjestelmien tuottajilla tai hyödyntäjillä ei välttämättä ole mahdollisuutta tai motivaatiota arvioida, miten eri teknologiat yhdessä ja erikseen vaikuttavat yksilöihin ja näiden toiminnan mahdollisuuksiin ja hyvän elämän edellytyksiin[219]. Kehitys vaikuttaa johtavan siihen, että haittojen ennalta estämisen sijaan joudumme yhä laajemmin varautumaan haittojen jälkikäteiseen minimointiin[220].

3.3.2 Seurannaishaitat

Seurannaishaitat etääntyvät teknologian välittömästä toimintapiiristä eli ne ilmenevät ajallisesti tai maantieteellisesti erillään haittaan johtavasta algoritmisesta toiminnasta. Seurannaishaitat kytkeytyvät tekoälyteknologioiden transformatiivisiin ominaisuuksiin: kuten muutkin teknologiat aiemmin, myös tekoälyteknologiat muuttavat ympäristöään ja käyttäjiään, ja vaikutukset leviävät väistämättä ajallisesti ja

217 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1). 188.

218 Ibid. 189.

219 Esimerkiksi Hollannissa SyRI-algoritmi keräsi tietoa useista datalähteistä tunnistaakseen sosiaalietuuksien väärinkäyttäjät, mikä johti vakaviin haittoihin. Ks. Chiusi, F., Fischer, S., Kayser-Bril, N., & Spielkamp, N. (2020). Automating society report 2020. AlgorithmWatch. 160–172.

220 Ks. myös esimerkiksi Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *Computer Law & Security Review*, *44*, 105657.

maantieteellisesti laajalle alueelle. Tässä alaluvussa pääpaino onkin uusissa, algoritmisten järjestelmien mahdollistamissa mekanismeissa, joiden kautta algoritmisia, yksiöön kohdistuvia seurannaishaittoja voi ilmetä. Koska seurannaishaitat eivät kiinnity teknologian välittömään toimintaan, niiden taustalla vaikuttavat teknologian lisäksi myös monet muut tekijät, joita käyttöympäristö ja käytön tavat määrittävät. Näin ollen seurannaishaittojen tarkastelussa olennaiseksi nousee yhtäältä, millaisissa yhteyksissä algoritmisia seurannaishaittoja syntyy eli miten ja millaiset teknologioiden käyttötavat ja -kontekstit vaikuttavat haittapotentiaaliin, ja toisaalta, minkälaisiin haittoihin tämä voi johtaa.

3.3.2.1 Heijastusvaikutukset

Ensimmäisessä osatutkimuksessa[221] yhtenä esimerkkitapauksena arvioitiin Yhdysvaltojen Michiganin osavaltiossa käytössä olleen MiDAS-petoksentunnistusalgoritmin vaikutuksia yhteiskunnallisten haittojen näkökulmasta. Järjestelmän tekemät virheelliset päätökset työttömyyskorvaushakemusten käsittelyssä veivät perusteetta monilta tulot ja leimasivat nämä potentiaalisiksi rikollisiksi. Kyse on välittömistä, yksilöön kohdistuvista haitoista. Haitat eivät kuitenkaan jääneet tähän. Tulojen menetys heijasti vaikutuksensa esimerkiksi lainanmaksukykyyn, minkä seurauksena monet menettivät asuntonsa. Mahdollisuudet osallistua yhteiskunnan toimintaan, toteuttaa itseään ja elää yhteisön jäsenenä heikkenivät paitsi taloudellisen tilanteen romahtamisen seurauksena myös siksi, että järjestelmän virheellisestä toiminnasta seuranneet petossyytteet leimasivat hakijat perusteetta epärehellisiksi ja rikollisiksi. Kun ihmisen henkilökohtainen tilanne muuttuu äkisti radikaalisti huonompaan, sillä on todennäköisesti vaikutusta myös tämän mielenterveyteen ja psyykkiseen kantokykyyn[222]. Kaikki tämä vaikuttaa yksilön toimintakykyyn.

Virheellisten päätösten seuraukset näkyvät myös yhteiskunnan toiminnassa. Mikäli automaatio lisää virheitä tai epäilyjä virheistä, tarve myös oikaisuvaatimuksille[223] tai muutoksenhakuprosesseille väistämättä kasvaa. Tällöin automaatio voi kyllä nopeuttaa ensivaiheen päätöksentekoa, mutta samalla siirtää ongelmia seuraa-

221 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

222 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1). 186.

223 Suomessa hallintolaki edellyttää, että automaattisen päätöksenteon kohteilla on aina maksuton mahdollisuus oikaisuvaatimuksen tekemiseen. Ks. Hallintolaki 8b luku 53f § 1 momentti.

valle tasolle[224]. Yksilöille tämä tietää aikaa vieviä ja mahdollisesti raskaita muutoksenhakuprosesseja, joiden aikana väärän päätöksen aiheuttamat haitat voivat heijastaa vaikutuksensa yhä laajempaan ympäristöön. On todennäköistä, että virheellisten päätösten mukanaan tuomat haitat kohdentuvat voimakkaimmin niihin, joiden asema yhteiskunnassa on jo valmiiksi heikko[225].

Seurannaishaitat eivät kuitenkaan aina perustu algoritmisen järjestelmän toiminnan virheisiin tai vääristymiin. Algoritmiset teknologiat vaikuttavat perustavanlaatuisilla tavoilla yhteiskunnan toimintoihin, sosioekonoteknisiin järjestelmiin sekä ihmisen ajatteluun ja toimintaan[226]. Muutos voi olla vähittäinen, ja se voi syntyä useampien teknologioiden yhteisvaikutuksena. Seurannaishaittoina arvioitavaksi nousevat paitsi yksittäiset teknologiat ja niiden toiminnan haitalliset seuraukset, myös teknologioiden aikaansaamista neutraaleista tai positiivisista muutoksista seuraavat haitalliset kehityskulut.

3.3.2.2 Haittojen tuottamisen tapojen muutokset

Algoritmiset järjestelmät lisäävät mahdollisuuksia tarkoituksellisesti *tuottaa* haittoja. Käytännössä tämä tarkoittaa, että algoritmiset järjestelmät voidaan valjastaa pahantahtoisten tarkoitusperien toteuttamiseen. Kun haitalliseksi määritellään toiminta, joka heikentää mahdollisuuksia hyvään elämään ja inhimilliseen kukoistukseen, erilaiset ihmisen autonomiaa ja toimintakykyä heikentävät teknologiat näyttäytyvät vääjäämättä haitallisina. Tällaiset haitat ovat pääsääntöisesti seurannaishaittoja: ne eivät ilmene algoritmisen teknologian välittömänä seurauksena vaan ennemmin vähittäin ajan kuluessa ja vaikutusten kerrostuessa. Algoritmisesti kohdennettu mainonta ja personoitu uutissyöte[227] niin sosiaalisen kuin perinteisen median alustoilla lisäävät mahdollisuuksia paitsi manipulointiin[228] myös disinfor-

224 Ks. myös Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

225 Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *New York, NY: AI Now Institute*.

226 Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

227 Kohdennettu mainonta mahdollistaa myös syrjinnän, mikä voi entisestään heikentää vähemmistöjen asemaa, ks. esimerkiksi Angwin, J., Tobin, A., & Varner, M. (2017). Facebook (still) letting housing advertisers exclude users by race. (21.11.2017) *ProPublica*.

228 Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Computational Propaganda Research Project*;

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, *48*(2), 204–231.

maation[229] ja vihapuheen levittämiseen[230]. Internetin informaatiotulva myös vaikeuttaa faktojen tarkastamista ja manipuloinnin tunnistamista entisestään. Haitan yhteys algoritmiseen teknologiaan on kuitenkin välillinen: haittoja syntyy vasta, kun esimerkiksi manipuloivat teknologiat *vaikuttavat* käyttäjiensä ajatteluun, tietämykseen ja ymmärrykseen siten, että näiden mahdollisuus autonomiseen toimintaan heikkenee. Esimerkkejä on lukemattomia, ja niistä voidaan mainita haitat, joita pahantahtoinen tuuppiminen (*sludge*[231]) tai algoritmien hyödyntäminen epäasialliseen tiedonkeruuseen ja valvontaan[232] aiheuttavat ihmisten autonomialle ja vapaudelle.

Sen lisäksi, että algoritmeja on mahdollista suunnitella ja tuottaa haitallisia tarkoitusperiä varten, myös olemassa olevia algoritmeja on joissain tilanteissa mahdollista manipuloida[233] siten, että niiden toiminta muuttuu haitalliseen suuntaan. Algoritmien manipulointi vaatii tietoa. Manipuloijan tulee tietää teknologian toiminnan tavoite ja ymmärtää riittävällä tavalla sen toiminnan logiikkaa. Tällöin algoritmin tunnettuja ominaisuuksia on mahdollista hyödyntää tavoilla, jotka saavat algoritmin toimimaan manipuloijan toivomalla, suunnitellusta poikkeavalla tavalla[234]. Osa manipulointipyrkimyksistä lienee sinänsä neutraaleja, eikä pyrkimystä haitan aiheuttamiseen ole. Tämä ei kuitenkaan tarkoita, etteikö haittaa voisi syntyä joko yllättävänä ja odottamattomana seurauksena tai tavoiteltuun lopputulokseen kytkeytyvänä väistämättömänä oheisvaikutuksena.

Joitakin algoritmeihin kohdistuvia manipulointitapauksia on tunnistettu. Esimerkiksi pörssikaupan markkina-algoritmeja[235] ja sosiaalisen median suosittelualgoritmeja[236] on onnistuttu ohjailemaan kyseenalaisin keinoin toimimaan niiden käyttötarkoitusta vastaamattomalla tavalla[237]. Sosiaalisen median algoritmeja voi manipu-

229 Ks. Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97*, 202–215.

230 Horwitz, J., & Seetharaman, D. (2020). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. (26.5.2020). *The Wall Street Journal*.

231 Thaler, R. H. (2018). Nudge, not sludge. *Science*, *361*(6401), 431–431.

232 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

233 Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 225.

234 Ibid.

235 Lin, T. C. (2016). The new market manipulation. *Emory LJ*, *66*, 1253.

236 Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 225.

237 Pörssikaupan ja sosiaalisen median vaikutukset voivat myös kumuloitua kiinnostavasti, kuten Redditin r/wallstreetbets-yhteisön vaikutukset osakemarkkinoihin osoittaa. Ks. esimerkiksi Jones, H., & Hietanen, J. (2023). The r/wallstreetbets 'war machine': Explicating dynamics of consumer resistance and capture. *Marketing Theory*, *23*(2), 225–247.

loida, koska tiedetään, että suosittelualgoritmit määrittävät julkaisujen näkyvyyttä muun muassa niiden saamien reaktioiden, kommenttien ja jakojen perusteella. Jos reaktioita lisätään keinotekoisesti, algoritmit päättelevät, että tietty julkaisu nauttii suurta suosiota, minkä vuoksi sitä kannattaa näyttää yhä useammalle taholle. Tällä on vaikutusta ihmisiin. Ihmisillä on taipumus mukautua vallitsevaan mielipiteeseen[238], jolloin mielipiteiden ja tiedon näkyvyyteen ja näennäiseen suosioon vaikuttamalla on mahdollista muokata yksittäisten henkilöiden näkemyksiä. Tätä voidaan hyödyntää esimerkiksi poliittisessa vaikutustyössä ja markkinoinnissa (*online astroturfing*[239]), ja mielipidemuokkauksen kautta on mahdollista vaikuttaa myös esimerkiksi julkiseen keskustelukulttuuriin[240]. Mielipidemuokkaaminen ja poliittiset vaikutuspyrkimykset heikentävät jälleen yksilön autonomiaa ja voivat altistaa myös etenkin mielenterveyteen kohdistuville haitoille.

3.3.2.3 Teknologiaperusteinen syrjäytyminen ja pahoinvointi

Ihmisten mahdollisuudet ja kyvyt yhtäältä hyödyntää algoritmisia teknologioita ja toisaalta tunnistaa ja välttää niiden haittapotentiaalin realisoituminen omassa elämässään vaihtelevat[241]. Laaja-alainen ja yhä kiihtyvä algoritmisaatio kasvattaa teknologisen osaamisen ja ymmärryksen merkitystä entisestään. Teknologisen kehityksen vanavedessä muuttuu myös käsitys siitä, minkälainen tieto on yhteiskunnassa niin työ- kuin arkielämässä ylipäätään merkityksellistä[242]. Tämä ilmiö on osa modernin yhteiskunnan eritahdistumista. Rosa nostaa esille sukupolvien kuilun (*generation break*): viisaus ei enää kiinnity iän tuomaan ymmärrykseen, vaan sen sijaan ikääntymiseen liitetään yhä voimakkaammin ajatus siitä, että sen myötä kyky selvitä muuttuvassa maailmassa heikkenee voimakkaasti[243].

Eri sukupolvien lisäksi vastaava kuilu vaikuttaa myös saman ikäluokan sisällä. Van Dijk käyttää käsitettä digitaalinen jakolinja (*digital divide*)[244] puhuessaan digi-

238 Ilmiö tunnetaan englanniksi termillä *conformity*. Ks. esimerkiksi Sunstein, C. R. (2019). *Conformity*. New York University Press.

239 Katso esimerkiksi Chan, J. (2022). Online astroturfing: A problem beyond disinformation. *Philosophy & Social Criticism*, 01914537221108467.

240 Ks. myös esimerkiksi Googlen suosittelualgoritmiin liittyen: Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, communication & society*, *20*(1), 63–80.

241 Esimerkiksi uusien teknologioiden vaikutuksista terveydenhuollon yhdenvertaisuuteen on puhuttu jo pitkään. Ks. Glied, S., & Lleras-Muney, A. (2008). Technological innovation and inequality in health. *Demography*, *45*, 741–761.

242 Myös Jasanoff, S. (2004). The idiom of co-production. Teoksessa *States of knowledge* (1–12). Routledge.

243 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 116.

244 Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

taalisten teknologioiden aiheuttamasta eriarvoistumisesta. Sopeutuminen ympäristöön, jonka toimintoja määrittävät algoritmiset teknologiat, vaatii yksilöiltä riittäviä resursseja, ja tällaisia resursseja ei ole jokaisella. Tämä luo tilanteen, jossa algoritmisaation aikaansaamista muutoksista juontuvat hyödyt ja toisaalta myös haitat kohdistuvat eri tavoin eri ihmisiin. Vaatimus teknologisesta osaamisesta heikentää väistämättä joidenkin ihmisten mahdollisuuksia yhteiskunnassa. Tätä ilmiötä kuvataan termillä digitaalinen tai teknologinen ekskluusio. Pääsy teknologioiden pariin ja teknologinen osaaminen vaikuttavat siihen, minkälaisia mahdollisuuksia yksilöllä on niin taloudelliseen, sosiaaliseen ja poliittiseen toimintaan kuin kulttuuri- ja erilaisten institutionaalisten palveluiden hyödyntämiseen[245]. Mikäli teknologisia valmiuksia ei ole, mahdollisuudet toimia yhteiskunnassa heikkenevät.

Digitaalisen ekskluusion ja siitä seuraavien haittojen voi arvioida lisääntyvän, jos ja kun palveluiden saatavuus on yhä voimakkaammin kiinni mahdollisuuksista ja kyvystä toimia nopeasti muuttuvassa algoritmisessa ympäristössä. Etenkin, kun ihmisten tekemä työ väistyy halvemman automaation tieltä[246], on riskinä, että yhä harvemmissa tilanteissa apua saadakseen voi kääntyä ihmisen puoleen. Tämä asettaa vaikeaan asemaan henkilöt, jotka eivät syystä tai toisesta pysty tai osaa toimia algoritmisessa ympäristössä: esimerkiksi monet iäkkäät, vammaiset, kielitaidottomat ja vähäosaiset henkilöt. Tämä heikentänee varsinkin jo valmiiksi heikossa asemassa olevien ihmisten asemaa yhteiskunnassa, lisää eriarvoisuutta ja haastaa ihmisten toiminnan mahdollisuuksia[247].

Teknologinen ekskluusio ja ihmisryhmien ja sukupolvien välinen kuilu todennäköisesti syvenevät entisestään, kun yhteiskunnallinen muutos ja siihen liittyvä algoritmisaatio kiihtyvät. Yhä nopeammin muuttuvassa yhteiskunnassa ihmisen hankkima tieto ja osaaminen ovat yhä lyhyemmän aikaa merkityksellisiä. Historian merkitys tulevaisuuden määrittäjinä heikkenee[248], ja yhä tärkeämmäksi muodostuu kyky omaksua uutta tietoa ja opetella uusia taitoja kiihtyvällä tahdilla. Nopea yhteiskunnallinen muutos vaatii yksilöiltä joustavuutta ja kykyä jatkuvaan mukautumiseen. Mikäli yksilö ei syystä tai toisesta pysty mukautumaan muutokseen, hänen kykynsä toimia yhteiskunnassa rajoittuu[249]. Jos ja kun voidaan pitää todennäköisenä, että kiih-

245 Czaja, I., & Urbaniec, M. (2019). Digital exclusion in the labour market in European countries: Causes and consequences. *European Journal of Sustainable Development*, *8*(5), 324–336. 333.

246 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research. 25.

247 Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E., & Winfield, A. (2020). *The ethics of artificial intelligence: Issues and initiatives*. European Parliamentary Research Service.

248 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 238.

249 Ibid. 243–244.

tyvä algoritmisaatio lisää teknologista ekskluusiota ja yhä suurempi osuus ihmisistä rajautuu vähintään joidenkin yhteiskunnan toimintojen ulkopuolelle tai jää vaille tarvittuja palveluita, on jokseenkin varmaa, että algoritmisaatioon kiinnittyvät haitat lisääntyvät[250]. Yksilötasolla ne voivat kohdistua niin terveyteen, autonomiaan, suhteelliseen asemaan kuin myös taloudelliseen turvallisuuteen.

Toisaalta myös ne, jotka eivät joudu kärsimään digitaalisesta ekskluusiosta, voivat kohdata uusia ongelmia. Jo perinteiset digitaaliset teknologiat ja niiden käyttäminen voivat lisätä muun muassa turvattomuutta ja riskiä joutua rikoksen tai hyväksikäytön uhriksi. Riski kasvanee, kun digitaalisten teknologioiden käyttö lisääntyy ja teknologiat muuttuvat algoritmisaation myötä käyttäjälleen vaikeammin ymmärrettäviksi ja hallittavaksi, ja algoritmisia teknologioita käytetään yhä uusilla tavoilla pahantahtoisiin tarkoituksiin[251]. Erilaisilla teknologioilla on myös potentiaalia aiheuttaa addiktiota ja terveydellisiä haittoja[252]. Etenkin sosiaalisen median vaikutuksia on tutkittu viime vuosina paljon, ja on esitetty, että sosiaalisen median käyttäminen voi lisätä esimerkiksi mielenterveyden ongelmia[253]. Algoritmit vaikuttavat sosiaalisen median alustoilla suosittelemalla ja rajaamalla näytettävää sisältöä eli ne määrittävät, minkälaista sisältöä ihminen näkee. Se vaikuttaa ihmisen toimintavaihtoehtoihin[254]. Sosiaalinen media ei kuitenkaan typisty algoritmeihin, vaan se toimii alustana sosiaaliselle kanssakäymiselle ja toiminnalle. Sosiaaliseen toimintaan itsessään taas liittyy myös negatiivisia seikkoja, joilla on vaikutusta mielenterveyteen: muiden muassa epäterve vertailu, kiusaaminen ja häirintä[255]. On kuitenkin selvää, etteivät algoritmit ole merkityksettömiä myöskään mielenterveysongelmien ja sosiaalisen median suhdetta tutkittaessa.

3.3.2.4 Vaikutukset työhön

Algoritmisaatio vaikuttaa myös työmarkkinoihin. Teknologisen kehityksen myötä kasvava määrä työtehtäviä on siirtynyt koneille, ja toisaalta uusia ammattiryhmiä on

250 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research.

251 Ks. esimerkiksi Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access.

252 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research.

253 Ks. esimerkiksi Karim, F., Oyewande, A. A., Abdalla, L. F., Ehsanullah, R. C., & Khan, S. (2020). Social media use and its connection to mental health: a systematic review. *Cureus*, *12*(6).

254 Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The information society*, *16*(3), 169–185.

255 Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

syntynyt[256]. On argumentoitu, että siinä missä digitalisaatio on jo aiheuttanut työttömyyttä ja painanut palkkoja alaspäin, tekoälyn hyödyntämisen yleistyminen voi jatkaa ja voimistaa samaa kehitystä[257]. Aiemmin on uskottu, että koneet ja tekoäly pystyvät tekemään yksinkertaiset työt, joissa on paljon toistoa ja vain vähän tarvetta luovuudelle[258]. Generatiivisen eli tekstiä, kuvia tai ääntä tuottavien tekoälyteknologioiden kehittämisen myötä tilanne on muuttunut: uudet tekoälyteknologiat suoriutuvat jo monista luovista ja laajaa asiantuntemusta vaativista tehtävistä, usein jopa ihmistä paremmin. Uusien generatiivisten tekoälyteknologioiden vaikutuksia työllisyyteen on vielä vaikea arvioida, mutta todennäköistä on, että ne muuttavat työelämää merkittävällä tavalla ja mahdollistavat yhä laajemmin työtehtävien automaation ja ihmistyövoiman vähentämisen[259]. Myös vaatimukset teknologiselle osaamiselle ja medialukutaidolle kasvanevat, mikä voi syventää digitaalisesta ja algoritmisesta erottelusta johtuvaa kuilua: nykyään jo työn löytämisen mahdollisuus on pitkälti sidottu teknologisiin taitoihin, sillä avoimista työpaikoista ilmoitetaan ja niitä haetaan pitkälti digitaalisessa ympäristössä.

Algoritmisaatio voi siis lisätä työttömyyttä[260], minkä lisäksi se voi luoda epävarmuutta töiden jatkuvuudesta. Molemmat seikat voivat aiheuttaa haittaa taloudelliselle turvallisuudelle, sillä jo pelko töiden menettämisestä lisää turvattomuutta, mikä voi heikentää myös psyykkistä hyvinvointia[261]. Pitkään kestävä taloudellinen turvattomuus voi aiheuttaa haittaa terveyden lisäksi myös autonomialle sekä asemalle yhteiskunnassa ja suhteessa muihin.

Seurannaishaittoja voi siis syntyä paitsi siksi, että algoritmisaatio vaikuttaa sosioekonomisiin järjestelmiin, myös siksi, että se muokkaa ihmisten toimintaympäristöä ja toiminnan mahdollisuuksia ja vaatimuksia. Kuten yllä esitettiin, teknologinen ekskluusio ja siitä aiheutuvat haitat todennäköisesti lisääntyvät, mikäli henki-

256 Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: how technology changes labor demand. *Journal of Economic Perspectives*, *33*(2), 3–30.

257 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research. 23–24.

258 Esimerkiksi Kim, Y. J., Kim, K., & Lee, S. (2017). The rise of technological unemployment and its implications on the future macroeconomic landscape. *Futures*, *87*, 1–9. 8.

259 Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research. 26.

260 Mondal, S., Das, S., & Vrana, V. G. (2023). How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. *Technologies*, *11*(2), 44;
Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research.

261 Katso työttömyyteen liittyvästä turvattomuudesta Shoss, M. K. (2017). Job insecurity: An integrative review and agenda for future research. *Journal of management*, *43*(6), 1911–1939.

löllä on liian vähäiset mahdollisuudet, kyvyt tai halu hyödyntää uusia teknologioita. Samaan aikaan kuitenkin myös *liiallisen* digitaalisten järjestelmien hyödyntämisen on argumentoitu aiheuttavan haittoja[262]. Sama ilmiö vaikuttaisi jopa voimakkaammalta algoritmisten teknologioiden kohdalla.

3.3.2.5 Algoritmisten teknologioiden kertautuvat vaikutukset

Monessa eri kontekstissa toimivat ja mahdollisesti toisiinsa yhteydessä olevat algoritmit voimistavat toistensa vaikutuksia[263] ja vaikeuttavat vaikutuksesta irtaantumista. Etenkin tiedon rajautuminen tai sen vääristyminen vaikuttavat siihen, minkälaisen ymmärryksen pohjalta ihminen päätöksiään tekee. Jos tiedon rajautuminen johtaa laajamittaisesti siihen, että ihmisten käsitykset todesta ja todellisesta heikkenevät tai jopa murenevat, siirrytään totuudenjälkeiseen aikaan, jossa tietoa on vaikeaa, ellei mahdotonta, erottaa mielipiteistä[264]. Seurauksena ihmisten luottamus yhteiskunnallisiin instituutioihin, kuten vapaaseen mediaan ja tieteeseen, voi heikentyä. Heikko luottamus yhteiskuntaan taas tarjoaa entistä paremman mahdollisuuden vakuuttaa ihmiset vaihtoehtoisista mahdollisuuksista, jolloin manipulointi käy yhä helpommaksi ja yksilöiden mahdollisuudet tietoon pohjaavaan autonomiseen toimintaan heikkenevät entisestään.

Muutokset ihmisten tietämyksessä, ymmärryksessä ja arvoissa todennäköisesti muokkaavat vähittäin myös kollektiivisia moraalisia arvoja, sivistystä, sosiaalisia normeja, kulttuuria sekä yhteiskunnallisia instituutioita ja rakenteita. Yhteiskunnalliset järjestelmät syntyvät ja muuntuvat inhimillisen toiminnan vaikutuksesta samalla, kun yhteiskunnalliset rakenteet vaikuttavat toimijoihin ja näiden toimintamahdollisuuksiin[265]. Algoritmisilla teknologioilla on yhteiskunnallisessa muutoksessa yhä suurempi rooli, sillä ne pystyvät vaikuttamaan niin rakenteisiin kuin toimijoihinkin tarjoamalla yhä uusia mahdollisuuksia toimintojen muutoksille ja niiden tehostamiselle. Tämä muutosvoima näkyy läpi yhteiskunnan. Kuten aiemmin on tuotu ilmi, palveluita ja päätöksentekoa yhteiskunnan eri sektoreilla automatisoidaan, työmarkkinoilla vaaditaan yhä enemmän algoritmista osaamista, ja niin valti-

262 Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons. Luku 7.

263 Personoitu käyttäjäkokemus perustuu vahvasti siihen, että ihmisen mieltymyksistä kerätään tietoa ja algoritmien avulla näkyville nostetaan käyttäytymishistorian perusteella optimoitua sisältöä. Kun useampi internet-sivusto tai sosiaalisen median alusta toimii samalla logiikalla, vaikutukset kumuloituvat ja kertautuvat. Ks. esimerkiksi Eg, R., Tønnesen, Ö. D., & Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, *9*, 100253.

264 Sismondo, S. (2017). Post-truth? *Social studies of science*, *47*(1), 3–6.

265 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration. Cambridge: Polity Press.

olliset toimijat kuin yrityksetkin hyödyntävät yhä laajemmin dataa, mikä on herättänyt keskustelua niin kutsutun valvontakapitalismin (*surveillance capitalism*) ongelmista[266]. Samalla internet ja algoritmiset toimintaympäristöt ovat kasvattaneet merkitystään arkisessa elämässä, mikä tarkoittaa, että algoritmiset teknologiat vaikuttavat yhä enemmän yksilöiden toimintamahdollisuuksiin, päätöksentekoon, sosiaaliseen kanssakäymiseen ja tiedonhakuun. Voidaan pitää jokseenkin todennäköisenä, että algoritmisaation vaikutukset yhtäältä yhteiskunnallisiin rakenteisiin ja toisaalta ihmisten toimintatapoihin ja -mahdollisuuksiin voimistavat toisiaan, mikä kiihdyttänee algoritmista transformaatiota entisestään[267]. Seuraukset voivat olla ennalta-arvaamattomia, sillä kiihtyvä vauhti vaikeuttaa muutosten hallintaa. Mahdolliset haitat voivat kohdistua paitsi yksilöihin, myös laajemmin yhteisöihin ja ryhmiin, joihin kohdistuvia haittoja käsittelen seuraavaksi.

3.4 Yhteisöihin ja ryhmiin kohdistuvat haitat

3.4.1 Välittömät haitat

Yhteisöihin tai ryhmiin kohdistuvat haitat ovat yhteisön tai ryhmän jäsenten kokemien haittojen summa silloin, kun niiden syntymistä määrittää ryhmään kuuluminen. Haitat kohdistuvat lähtökohtaisesti yksilöihin, mutta ne voivat johtua tiettyyn ryhmään kuulumisesta eli esimerkiksi ryhmän jäsenille ominaisten seikkojen tulkitsemisesta negatiivisella tavalla siten, että se johtaa haitalliseen lopputulokseen. Tutkimukset osoittavat, että esimerkiksi koneoppivin menetelmin toteutetut päätöksentekojärjestelmät toistavat herkästi yhteiskunnan rakenteellisia vinoumia ja asettavat joihinkin ryhmiin kuuluvat epäedullisempaan asemaan kuin valtaväestöön kuuluvat[268]. Mikäli näin tapahtuu, seurauksena ilmenevät virheelliset päätökset eivät kohdistu satunnaisiin yksilöihin vaan kokonaisiin ryhmiin. Teknologioiden systemaattinen toiminta[269] levittää mahdollisesti seuraavat haitat väistämättä jokaiseen ryhmän jäseneen, joka kyseisen teknologian kanssa päätyy tekemisiin.

266 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.

267 Myös Rosa argumentoi, että yhteiskunnan eri osa-alueiden muutos kiihdyttää muiden osa-alueiden muutosta. Ks. Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 194.

268 Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *AI Now Institute*.

269 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1). 192.

3.4.1.1 Algoritminen syrjintä

Yhä selvemmäksi on käynyt, että tekoälyteknologiat vaikuttavat eri ryhmiin eri tavoin, ja että haitat kohdistuvat usein voimakkaammin jo valmiiksi marginalisoituihin ryhmiin[270]. Tekoälyn kehittäminen on pitkälti muutamien suurten, monikansallisten yritysten käsissä, ja vain harvalla ihmisellä on todellinen mahdollisuus vaikuttaa kehityksen suuntaan. Näin ollen yritysten päätökset vaikuttavat valtavasti siihen, missä yhteyksissä erilaisia tekoälyjärjestelmiä otetaan käyttöön, mitä arvoja ne edustavat tai minkälaista maailmaa ne rakentavat[271]. Samaan aikaan modernien yhteiskuntien jäsenten on mahdotonta välttää tekoälyteknologioiden vaikutuksia elämässään, mikä heikentää paitsi yksilöiden, myös yhteisöjen ja ryhmien mahdollisuuksia omalla toiminnallaan välttää algoritmisten teknologioiden heihin kohdistamia haittoja tai ylipäätään kontrolloida elinympäristöään.

Feministisen teorian yksi monista saavutuksista on intersektionaalisuuden käsitteen nostaminen yhteiskunnalliseen keskusteluun. Intersektionaalisuus tarkoittaa pyrkimystä huomioida riski moniperustaiselle syrjinnälle, ja se soveltuu myös algoritmisten teknologioiden kontekstiin[272]. Rakenteellista syrjintää joutuvat kohtaamaan monet eri ryhmät. Henkilöt, jotka kuuluvat samanaikaisesti useampiin marginalisoituihin ryhmiin, voivat kohdata risteämien takia syrjintää, joka on joko voimakkaampaa tai erilaista kuin yksittäiseen marginalisoituun ryhmään kohdistuva syrjintä. Kiellettyjen syrjintäperusteiden lista on lainsäädännössä määritelty. Vaikka algoritmiselta järjestelmältä estettäisiin laissa kiellettyjen, syrjivien ominaisuuksien hyödyntäminen päätöksenteossa, voivat etenkin syviin neuroverkkoihin pohjautuvat järjestelmät oppia niin kutsuttuja sijaismuuttujia, joiden kautta kielletty syrjintäperuste siirtyy järjestelmän toimintaan[273]. On vaikea tunnistaa, mitä muuttujia itsenäisesti toimivat, monimutkaiset, selittämättömät järjestelmät käyttävät. Tämä on haaste yksittäistenkin ryhmien kohdalla, ja moniperustaisen syrjinnän kohdalla ongelmallisen logiikan tunnistaminen on vielä vaikeampaa. Tämä lisää riskiä sille, että rakenteellisten vinoumien siirtymistä algoritmiseen päätöksentekoon ei huomata.

270 Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *AI Now Institute*.

271 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.

272 Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. Konferenssijulkaisu, *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 485–494).

273 Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Computer Law & Security Review, 41.

3.4.1.2 Muuttuvat käytännöt

Vaikka tekoälyjärjestelmä ei olisi viimesijainen päätöksentekijä, on todettu, että ihmiset luottavat järjestelmien neutraaliuteen usein liiaksikin[274]. Tämä voi lisätä entisestään syrjiviä käsityksiä ja käytänteitä. Yksi tunnettu esimerkki algoritmien potentiaalista toistaa yhteiskunnan vinoumia ja sitä kautta päätyä tiettyjä ryhmiä syrjiviin lopputuloksiin on Yhdysvaltojen tuomioistuimissa laajasti käytetty Compas-algoritmi. Se ennusti rikosprosessissa rikoksentekijän uusimistodennäköisyyttä ja arvioi systemaattisesti tummaihoisten rikoksentekijöiden uusimisriskin valkoihoisia korkeammaksi. Arvioitu riskin suuruus määräytyi rikoksenuusimisista saadun datan perusteella. Käytännössä algoritmi laski, kuinka todennäköisesti erilaiset rikoksenuusijat ovat datan keräämisalueella historiallisesti arvioiden jäänyt uusimisrikoksestaan kiinni, ja perusti tuloksensa siihen. Kun valvontaa on historiallisesti kohdistettu ja edelleen kohdistetaan rasistisin perustein rodullistettuihin, myös näiden rikoksista kiinnijäämisen todennäköisyys ja siten datan tunnistama uusimistodennäköisyys näyttäytyivät luonnollisesti korkeampina kuin valkoisilla rikoksentekijöillä. Tilanne on jokseenkin tyypillinen esimerkki dataan liittyvistä mahdollisista ongelmista: data ei välttämättä todellisuudessa kerro sitä, mitä sen oletetaan kertovan (misguided evidence[275]). Compas-algoritmin tapauksessa syntyneet haitat kohdistuivat etenkin tummaihoisiin, joita järjestelmän seurauksena kohdeltiin rikosoikeusjärjestelmässä valkoihoisia ankarammin. Näin tapahtui siitäkin huolimatta, että algoritmi ei toiminut viimesijaisena päätöksentekijänä, vaan tuomarilla säilyi valta tehdä päätös algoritmin ehdotuksesta poiketen.[276] Algoritminen järjestelmä näyttäisi kuitenkin tosiasiallisesti vaikuttaneen tehtyihin päätöksiin ja jopa voimistaneen rasistisia käytäntöjä oikeuslaitoksessa.

Konteksti, johon algoritminen teknologia kiinnittyy, vaikuttaa siihen, minkälaisia haittoja syntyy. Välittömät yhteiskunnalliset haitat, joita ryhmien jäseniin kohdistuu ryhmään kuulumisen perusteella, ovat vastaavia kuin yksilöihin kohdistuvat haitat. Ne voivat johtua esimerkiksi opetusdatan vinoumista, järjestelmän virheellisestä toimintalogiikasta, koneellisen harkinnan jähmeydestä tai henkilötietojen keräämisen tai yhdistelemisen ongelmista – eli niiden syntymiseen vaikuttavat samat seikat kuin yksilöihinkin kohdistuvien välittömien haittojen syntyynkin. Ongelmat

274 Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge.

275 Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 2053951716679679.

276 Bahl, U., Topaz, C. M., Obermüller, L., Goldstein, S., & Sneirson, M. (2023). Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. *UCLA L. Rev. Discourse*, *71*, 246. Tutkimuksessa osoitetaan, että Compas-algoritmin hyödyntäminen kasvatti eroa valkoisten ja tummaihoisten kohtelussa rikosprosessissa.

kuitenkin voivat korostua ja niiden havaitseminen voi olla vaikeaa etenkin, mikäli eri vähemmistöjen intersektioita ja niihin liittyviä riskejä ei tunnisteta. Ongelmien seurauksena syntyvät haitat voivat vaikuttaa ryhmien jäsenten taloudelliseen tilanteeseen, terveyteen, autonomiaan ja asemaan. Kun haitat kohdistuvat yksilöihin siksi, että nämä kuuluvat tiettyyn ryhmään, haitat kiinnittyvät valtarakenteisiin, epätasa-arvoon ja sorron historiaan. Nämä vaikuttavat vahvasti myös seurannaishaittojen synnyssä.

3.4.2 Seurannaishaitat

Seurannaishaitat, jotka kohdistuvat kokonaisiin ryhmiin, voivat etääntyä haittoja aiheuttavasta algoritmisesta teknologiasta huomattavasti. Tämä johtuu siitä, että tällaisia haittoja syntyy erityisesti algoritmisten järjestelmien käyttämisestä seuraavien muutosten seurauksena. Toki teknologiseen ekskluusioon liittyvät seurannaishaitat, joita käsiteltiin aiemmin, kohdistuvat myös ryhmiin. Erityisesti yhteiskunnassa jo valmiiksi heikossa asemassa olevat ryhmät, kuten matalasti koulutetut[277], ikääntyneet tai valmiiksi marginalisoidut tai syrjäytyneet[278], kärsivät ekskluusion vaikutuksista[279]. Algoritmisaatio ei myöskään pääse *hyödyttämään* tasapuolisesti kaikkia ryhmiä, mikä voi entisestään lisätä ja syventää eriarvoisuutta[280]. Algoritmiset järjestelmät kuitenkin vaikuttavat eri ryhmien asemaan myös huomattavasti vaivihkaisemmin vaikuttamalla kulttuuriin, asenteisiin ja yhteiskunnallisiin normeihin. Seuraavaksi keskitynkin erityisesti kehityskulkuihin, joita algoritmisten teknologioiden on todennettu voimistavan, ja joiden seurauksena vaikuttaisi siltä, että erilaiset yhteiskunnalliset käytännöt tai ihmisten asenteet ovat kiristyneet tai kiristymässä siten, että joidenkin ryhmien kohtaamat haitat lisääntyvät tai saattavat lisääntyä. Lisäksi pyrin tarkastelemaan, mitkä seikat algoritmisissa järjestelmissä lisäävät tällaisten kehityskulkujen mahdollisuuksia.

277 Zarouali, B., Helberger, N., & De Vreese, C. H. (2021). Investigating algorithmic misconceptions in a media context: Source of a new digital divide? *Media and Communication*, *9*(4), 134–144.

278 Sanders, C. K., & Scanlon, E. (2021). The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Journal of Human Rights and Social Work*, *6*, 130–143.

279 Sanders, C. K., & Scanlon, E. (2021). The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Journal of Human Rights and Social Work*, *6*, 130–143.

280 Terveydenhuollon kontekstissa ks. esimerkiksi Timmermans, S., & Kaufman, R. (2020). Technologies and health inequities. *Annual Review of Sociology*, *46*, 583–602.

3.4.2.1 Algoritmisten järjestelmien vaikutukset ajatteluun

Kuten yksilöihin kohdistuvia seurannaishaittoja käsittelevässä alaluvussa esitin, algoritmiset työkalut vaikuttavat ihmisten ajatteluun ja toimintaan. Niillä on vaikutusta siihen, mihin huomio kohdistetaan ja mitä seikkoja yksilö päätöksenteossaan painottaa[281]. On mahdollista, että algoritmiset teknologiat paitsi toistavat yhteiskunnan vinoumia, kuten rakenteellista rasismia, myös kiristävät ihmisten suhtautumista vähemmistöihin. Tällainen epäsuora vaikutus voi syntyä, mikäli luottamus teknologioiden neutraaliuteen johtaa kehäpäätelmään: jos ihminen tulkitsee yhteiskunnan vinoumia toistavat algoritmiset teknologiat neutraaleiksi ja oikeudenmukaisiksi, teknologioiden tekemät vinoutuneet päätökset perustelevat ja oikeuttavat näkemyksiä, jotka toistavat ja vahvistavat yhteiskunnan rakenteellista epäoikeudenmukaisuutta[282]. Vaikutukset voivat olla kauaskantoiset: Rakenteellinen sorto voi voimistua, ja sorron lisääntyessä luottamus paitsi teknologioihin myös instituutioihin, joissa niitä hyödynnetään, voi heiketä. Luottamuspula taas voi vähentää halua tukeutua yhteiskunnan instituutioihin[283], lisätä ääriajattelua[284] ja ruokkia vaihtoehtoisten järjestelmien kannatusta[285].

Algoritmisten teknologioiden vaikutuksia erilaisiin yhteiskunnallisiin ilmiöihin tai kehityskulkuihin voi olla vaikeaa tunnistaa tai näyttää toteen etenkin silloin, kun algoritmiset teknologiat toimivat näkymättömissä[286] ja vaikuttavat useisiin eri prosesseihin. Samoin on vaikeaa tunnistaa, määrittää ja mitata teknologian välittömästä vaikutuspiiristä etääntyviä seurannaishaittoja, jotka eivät ole helposti palautettavissa teknologian välittömään toimintaan. Mitä kauempana teknologiasta seurannaishaitat ilmenevät, niin ajallisesti kuin maantieteellisesti, sitä useampi tekijä todennäköisesti vaikuttaa haitan taustalla. Esimerkiksi ääriajattelun lisääntymiseen vaikuttanevat monimutkaiset sosioekonotekniset prosessit. Algoritmisilla teknologioilla on vaiku-

281 Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–33.

282 Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278.

283 Van Prooijen, J. W., Spadaro, G., & Wang, H. (2022). Suspicion of institutions: How distrust and conspiracy theories deteriorate social relationships. *Current opinion in psychology*, *43*, 65–69.

284 Van den Bos, K. (2020). Unfairness and radicalization. *Annual review of psychology*, *71*, 563–588.

285 Mondon, A., & Winter, A. (2020). Reactionary democracy: How racism and the populist far right became mainstream. Verso Books.

286 Elliott, A. (2019). The culture of AI: Everyday life and the digital revolution. Routledge.

tusta, sillä ne tarjoavat monille prosessille entistä parempia ja nykyään mahdollisesti myös välttämättömiä työkaluja ja toiminta-alustoja.

Ihmiset toimivat yhä useammin ympäristössä, jossa sekä toiminnan mahdollisuudet että ympäristö muovautuvat algoritmisten teknologioiden vaikutuksesta[287]. Internetin hakukoneiden ja sosiaalisen median suosittelualgoritmit rajaavat, mihin tietoon ihminen pääsee käsiksi, ja siten ne muokkaavat ihmisen käsitystä maailmasta[288]. Muuttunut käsitys maailmasta vaikuttaa ensi sijassa suoraan ihmisen toimintaan. Jos muutos koskettaa riittävän suurta osaa yhteiskunnan jäsenistä, se voi vaikuttaa myös yhteiskunnassa vallitseviin sosiaalisiin normeihin ja lopulta yhteiskunnan toimintaan. Mahdollisesti syntyvien haittojen taustalla voi siis olla monimutkaisia tapahtumaketjuja, kun teknologia, ympäristö ja inhimillinen toimija vaikuttavat toisiinsa moninaisilla ja toisinaan myös yllättävillä tavoilla. Etenkin, jos useammassa prosessissa tapahtuu muutos kohti haitallisempaa, voitaneen pitää todennäköisenä, että seurannaishaitat kertautuvat, syventyvät ja leviävät koskemaan laajasti kokonaisia ryhmiä. Samaan aikaan haittojen syntymekaniikan sirpaloituminen vaikeuttaa ensinnäkin haittoja aiheuttavien prosessien ja toisekseen niihin vaikuttavien algoritmisten teknologioiden tunnistamista.

Toisaalta ei ole vääjäämätöntä, että algoritmisten teknologioiden hyödyntäminen esimerkiksi hakukoneissa tai sisällön suosittelussa johtaa haittoihin. Hakukoneet ja suosittelualgoritmit, kuten muutkin tekoälyjärjestelmät, toimivat ohjelmointinsa mukaan: niiden toiminnan tavoitteet ovat ihmisten määrittelemiä. Näin ollen toimintamekanismeihin on mahdollista myös puuttua, mikä tarkoittaa, että esimerkiksi tiedon rajautumiseen tai vääristymiseen liittyviä ongelmia on mahdollista vähentää. Tämä voi tosin olla vaikeaa. Suurten teknologiayritysten intressissä on kuluttajien sitouttaminen sovellusten aktiiviseen käyttämiseen. Tiedossa on, että ainakin sosiaalisen median sovelluksiin sitouttaminen onnistuu tehokkaasti, kun niiden suosittelualgoritmit tarjoavat tunteita herättävää – ja usein vastakkainasetteluja lisäävää – sisältöä[289].

3.4.2.2 Tekoälyn tuotokset ja maailmankuvan muutokset

Algoritmisia teknologioita hyödynnetään paitsi päätöksenteossa ja olemassa olevan tiedon ja sisällön suosittelussa, rajaamisessa ja levittämisessä, myös taiteen, tieteen

[287] Muiden muassa Brownsword on arvioinut riskejä, joita liittyy teknologioiden avulla toiminnan sääntelyyn (*technological management*). Ks. Brownsword, R. (2019). *Law, technology and society: reimagining the regulatory environment*. Routledge.

[288] Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The information society*, *16*(3), 169–185.

[289] Horwitz, J. & Seetharaman, D. (2000). Facebook executives shut down efforts to make the site less divisive. (25.5.2000). *The Wall Street Journal*.

ja kulttuurin *tuottamisessa*. Näyttää todennäköiseltä, että tässäkin roolissa algoritmisten teknologioiden merkitys kasvaa entisestään tekstiä ja kuvia tuottavien generatiivisten tekoälyjärjestelmien yleistyessä.

Nykyään tekoälyjärjestelmien opetusdata on korostetun länsimaista, mikä tarkoittaa, että myös niiden tuottama sisältö kiinnittyy vahvasti länsimaiseen, valkoiseen kulttuuriperinteeseen[290]. Mikäli länsimaista kulttuuria toisintavat tekoälyjärjestelmät leviävät ympäri maailman, globaalit vaikutukset ovat väistämättä perustavanlaatuisia[291]. Kulttuurisen representaation kaventuminen tai yksipuolistuminen voi aiheuttaa haittoja yhä useampien ihmisten kulttuuriselle turvallisuudelle ja asemalle. Etenkin vähemmistöihin kuuluvien hyvinvointi voi heikentyä. Ennakkoluulojen, syrjinnän, vihan ja väkivallan uhan kanssa eläminen altistaa vähemmistöihin kuuluvat niin kutsutulle vähemmistöstressille[292] ja siitä seuraaville terveyshaitoille. Kuten aiemmin olen tuonut ilmi, algoritmisten teknologioiden on havaittu toistavan yhteiskunnassa esiintyvää rakenteellista syrjintää[293], mikä heikentää vähemmistöjen asemaa entisestään. Lisäksi, mikäli yksipuolistunut representaatio kaventaa jo ennestään kapeaa käsitystä tavanomaisesta tai sosiaalisesti hyväksytystä, on mahdollista, että yhä useampi inhimillinen piirre tulkitaan *poikkeavana*, mikä voi lisätä toiseuttavia (*othering*)[294] kehityskulkuja ja altistaa yhä useammat vähemmistöstressille.

Mikäli länsimainen tapa ymmärtää maailmaa leviää ja käy yhä hallitsevammaksi, siitä poikkeavat tavat nähdä, kokea ja ymmärtää voivat olla vaarassa heiketä tai jopa kadota, millä voi olla peruuttamattomia vaikutuksia tietoon, sivistykseen ja kulttuurien monimuotoisuuteen. Tiedon monipuolisuudella, sivistyksellä ja kulttuurien moninaisuudella on arvoa jo itsessään, ja niiden rapautuminen rajoittaa väistämättä mahdollisuuksia hyvään elämään – etenkin, kun hyvän elämän edellytyksenä hahmotetaan toimintamahdollisuuksien[295] kautta. Mikäli yhden ryhmän maailman-

290 Makhortykh, M., Urman, A., & Ulloa, R. (2021). Detecting race and gender bias in visual representation of AI on web search engines. *Advances in Bias and Fairness in Information Retrieval: Second International Workshop on Algorithmic Bias in Search and Recommendation, BIAS 2021, Lucca, Italy, April 1, 2021, Proceedings* (36–50). Cham: Springer International Publishing.

291 Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197.

292 Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological bulletin*, *129*(5), 674.

293 Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *AI Now Institute*.

294 Ks. toiseuttamisesta esimerkiksi Jensen, S. Q. (2011). Othering, identity formation and agency. *Qualitative Studies*, *2*(2), 63–78.

295 Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

katsomus syrjäyttää muut, muiden ryhmien mahdollisuudet toteuttaa perinteitään ja kulttuuriaan heikkenevät. Tämä aiheuttaa haittoja ryhmien jäsenten integriteetille[296]. Integriteetti kiinnittyy yhtäältä autonomiaan ja toisaalta yhteiskunnalliseen ja sosiaaliseen asemaan, ja näihin kohdistuvat haitat heikentävät vääjäämättä haittojen kohteen hyvinvointia. Teknologisen osaamisen ja tekoälyjärjestelmien kehittämisen keskittyminen voimakkaasti lähinnä Yhdysvaltoihin ja Kiinaan onkin herättänyt keskustelua nykymaailman kolonialismista[297], jossa teknologiamahdit määrittävät koko maailman kulttuuri- ja arvopohjan[298]. Tästä seuraavat haitat kohdistuvat myös laajasti yhteiskuntiin ja niiden toiminnan mahdollisuuksiin, joita käsittelen seuraavaksi.

3.5 Yhteiskuntaan kohdistuvat haitat

Yhteiskuntaan kohdistuvien haittojen arvioimista vaikeuttaa huomattavasti se, että yhteiskunta ei ole selkeärajainen, helposti määriteltävä entiteetti, jonka ihanteellinen tila olisi mahdollista yksiselitteisesti kuvata. Yhteiskunnat toimivat eri tavoin, jolloin myös erilaisia seikkoja pidetään toivottuina tai haitallisina. Näin ollen myös yhteiskuntaan kohdistuvien haittojen määrittely on vaikeaa: jos ihannetilaa ei ole mahdollista määritellä, teoriassa mitkä vain muutokset on mahdollista arvioida haitoiksi. Jotta yhteiskuntaan kohdistuvien haittojen arviointi olisi mielekästä, tässä työssä katson, että yhteiskunnan tehtävänä on tarjota kansalaisilleen puitteet hyvälle elämälle[299]. Näin ollen yhteiskuntaan kohdistuvina haittoina näyttäytyvät sellaiset muutokset, jotka heikentävät yhteiskunnan mahdollisuuksia tarjota kansalaisilleen edellytykset hyvälle elämälle. Hyvän elämän perusteita olen määritellyt ensimmäisessä luvussa.

Käytännössä yhteiskuntaan kohdistuvat haitat voivat horjuttaa yhteiskunnan välttämättömiä toimintoja ylläpitäviä rakenteita, esimerkiksi taloudellisia, oikeudellisia ja sosiaalisia instituutioita. Merkityksellisiä ovat myös arvot ja periaatteet, joita yhteiskunnassa ylläpidetään: demokraattisissa oikeusvaltioissa siis muiden muassa demokratia ja oikeusvaltioperiaate sekä yhteiskunnallinen vakaus, turvallisuus ja oikeudenmukaisuus, jotka oikeusvaltion voi nähdä mahdollistavan. Nämä ovat jokseenkin välttämättömiä edellytyksiä sille, että yhteiskunta pystyy järjestämään puitteet jäsentensä hyvälle elämälle. Käytännössä yhteiskuntaan kohdistuvien muutosten

296 Ibid.

297 Yle. (2023). Facebookin moderaattorit vaativat oikeuksiaan ja saivat potkut Keniassa – ”Facebook on uusi siirtomaavalta”. (10.5.2023).

298 Sahbaz, U. (2019). Artificial intelligence and the risk of new colonialism. *Horizons: Journal of International Relations and Sustainable Development*, (14), 58–71.

299 Näin asian näkee myös esimerkiksi Nussbaum. Ks. Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

haitallisuus lopulta määräytyy sen perusteella, miten ne vaikuttavat yhteiskunnan jäseniin. Yhteiskuntaan kohdistuvat haitat lisäävät yhteiskunnassa ilmeneviä, inhimillistä kukoistusta rajoittavia tai estäviä ilmiöitä heikentämällä yhteiskunnan mahdollisuuksia estää haittojen ilmenemistä tai muuttamalla yhteiskunnan käytänteitä siten, että haitat lisääntyvät. Yhteiskuntaa muuttavia tapahtumia tai kehityskulkuja, joiden seurauksena yksilöiden kohtaamat haitat eivät lisäänny, en sitä vastoin määrittele haitoiksi, vaikka ne muuttaisivatkin esimerkiksi oikeuslaitoksen toimintaa.

Pääosa yhteiskuntaan kohdistuvista algoritmisista haitoista on seurannaishaittoja. Algoritmisten teknologioiden välittömät vaikutukset kohdistuvat pääsääntöisesti ihmisiin, ja ihmisten kautta vaikutus siirtyy yhteiskunnalliselle tasolle, mikä väistämättä tarkoittaa, että yhteiskuntaan kohdistuvat vaikutukset etääntyvät teknologian toiminnasta niin paljon, että ne on perusteltua määritellä seurannaishaitoiksi. Kuitenkin, kuten ensimmäisessä osatutkimuksessa[300] esitetään, esimerkiksi pörssikauppaa käyvien algoritmien virheellinen toiminta voi pahimmillaan syöstä maailmantalouden globaaliin taloudelliseen taantumaan, jolloin välittömiä haittoja kohdistuu myös yhteiskuntien taloudelliseen järjestelmään ja sitä kautta yhteiskuntien toimintamahdollisuuksiin. Vastaavia välittömiä haitallisia vaikutuksia voi arvioida olevan myös muilla valtioiden ja maailman talousjärjestelmään kiinnittyvillä algoritmisilla teknologioilla.

3.5.1 Seurannaishaitat

Käsittelen tässä alaluvussa algoritmisiin teknologioihin linkittyviä kehityskulkuja, jotka muuttavat yhteiskuntaa haitallisempaan suuntaan. Vielä toistaiseksi aihetta ymmärretään ainoastaan rajallisesti. Tämä johtunee osaltaan siitä, että haitallisten muutosten tunnistaminen vaatii aikaa, sillä yhteiskunnalliset muutosprosessit ovat hitaita. Algoritmisten teknologioiden vaikutukset eivät ole tämän osalta poikkeus. Niiden vaikutukset kumuloituvat ja kertautuvat hitaasti yhteiskunnallisissa käytänteissä, ja muutosten laatu selviää aina vasta niiden vaikutusten kantautuessa takaisin yhteiskunnan jäsenten elämään.

Algoritmiset teknologiat vaikuttavat niin yhteiskuntaan ja sen instituutioihin kuin ihmisiin, kuten myös toimintaan, jota yhteiskunnan rakenteiden ja ihmisten vuorovaikutuksessa syntyy. Algoritmisten teknologioiden vaikutukset sirpaloituvat yhteiskunnassa ajallisesti ja maantieteellisesti laajalle alueelle. Haittojen taustalla vaikuttavien tekijöiden määritteleminen ja rajaaminen on vaikeaa, sillä yhteiskunnan rakenteet ja instituutiot ovat monimutkaisia kokonaisuuksia, jotka paitsi vaikuttavat

[300] Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

ympäristöönsä, myös muuntuvat ympäristönsä vaikutuksesta[301]. Algoritmiset teknologiat voivat vaikuttaa yhteiskunnan instituutioihin ja toimintoihin useiden vaikutussuhteiden kautta. Kuten aiemmin olen esittänyt, algoritmisten teknologioiden on uskottavasti argumentoitu muokkaavan esimerkiksi työelämää[302], päätöksentekoa[303], tiedonhakua ja -tuotantoa sekä käsitystä merkityksellisestä osaamisesta[304], sosiaalisen kanssakäymisen tapoja[305], ja myös yhteiskunnan kollektiivisia oikeudellisia, moraalisia ja kulttuurisia arvoja[306]. Erilaiset prosessit vaikuttavat yhteiskunnan toimintoihin eri tavoin.

Vaikka vaikuttavia tekijöitä ei olisi mahdollista täydellisesti erotella tai tunnistaa, on tärkeää ymmärtää, että algoritmiset teknologiat ovat äärimmäisen vaikuttavia teknologioita, ja niillä voi olla huomattavia haitallisia vaikutuksia myös yhteiskuntaan. Vaikutukset voivat kohdistua esimerkiksi talouden ja työelämän prosesseihin, yhteiskunnallisten instituutioiden toimintaan tai demokraattisten prosessien kannalta välttämättömään, yhteiskunnan jäsenten yleiseen tietotasoon siten, että yhteiskunta muuttuu aiempaa haitallisemmaksi. Ajan kuluessa tällaiset muutokset voivat johtaa siihen, että yhteiskunta köyhtyy ja menettää sen myötä toimintamahdollisuuksiaan, demokraattisten prosessien legitimiteetti heikkenee, luottamus yhteiskuntaan ja yhteiskunnassa vähenee ja erilaiset vaihtoehtoiset yhteiskuntamuodot kasvattavat kannatustaan. Näihin teemoihin liittyvät kehityskulut ovat tässä luvussa keskiössä.

3.5.1.1 Köyhtyvä valtio

Mikäli dystooppisimmat skenaariot automaation vaikutuksista työllisyyteen toteutuvat[307] ja tekoälyjärjestelmillä korvataan inhimillinen työnteko entistä laajemmin il-

301 Stones, R. (2017). *Structuration theory*. Bloomsbury Publishing.

302 Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: how technology changes labor demand. *Journal of Economic Perspectives*, *33*(2), 3–30.

303 Yeung, K. (2019). Why worry about decision-making by machine? In *Algorithmic regulation*. Oxford University Press.

304 Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., ... & Siemens, G. (2022). Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?. *Computers and Education: Artificial Intelligence*, *3*, 100056.

305 Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100.

306 Yeung, K. (2019). Responsibility and AI - A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe. 42.

307 Kuusi, O., & Heinonen, S. (2020). Tulevaisuuspolkuja kapeasta tekoälystä vahvaan tekoälyyn. *Tieteessä tapahtuu*, *38*(3). Skenaario 2 esittelee tilanteen, jossa tekoälyn kehitys johtaa yhteiskunnan sekasortoon.

man, että samalla turvataan ihmisten toimeentulo ja valtion tulovirta, on jokseenkin varmaa, että haittoja kohdistuu enenevissä määrin paitsi työttömyydestä kärsiviin, myös laajemmin yhteiskuntiin.

Kulutuksen jatkuvaan kasvuun perustuva talousjärjestelmä tarvitsee kansalaisia, jotka tekevät töitä ja siten mahdollistavat markkinatalouden toiminnan yhtäältä tuottamalla yhä enemmän tuotteita kulutettavaksi, toisaalta lisäämällä palkkatuloillaan omia mahdollisuuksiaan kuluttamiseen[308]. Valtio kerää tuloista ja kulutuksesta veroja, joilla yhteiskunnan ja sen instituutioiden toimintaa rahoitetaan. Mikäli algoritmisaatio lisää radikaalisti työttömyyttä, eikä yhteiskunnan jäsenten toimeentuloa onnistuta turvaamaan esimerkiksi radikaaleilla verotusjärjestelmän uudistuksilla[309] ja perustulolla[310], verotettavat tulot ja ostovoima heikkenevät, jolloin valtiolle kertyy väistämättä vähemmän verotuloja. On myös osoitettu, että jo pelko työttömyydestä vaikuttaa ihmisen toimintaan ja muun muassa kulutusvalintoihin[311]. Tämä itsessään voi jo vaikuttaa valtion verokertymään.

Riittävät hyvinvointipalvelut ovat hyvinvointivaltion keino taata mahdollisuudet hyvän elämän tavoittelun myös yhteiskunnan vähäosaisille. Valtion köyhtyessä yhteiskunnan tuottamien hyvinvointipalveluiden vähentäminen voi näyttää perustellulta. Tästä voi seurata kovia toimia menojen leikkaamiseksi, vaikka toimien vaikutuksista olisi ristiriitaisia arvioita[312], kuten Suomen nykyisen oikeistohallituksen pyrkimykset[313] osoittavat. Todennäköisesti monien mahdollisuudet elää hyvää elämää heikkenevät. Kun valtion hyvinvointipalvelut heikentyvät, se vaikuttanee suoraan yhdenvertaisuuden toteutumiseen lisäämällä eriarvoisuutta ja heikentämällä vähäosaisten mahdollisuuksia parantaa omaa asemaansa.

3.5.1.2 Tavanomaisen rajojen siirtyminen ja demokratia

Algoritmisaatio tarjoaa ihmisille ja yhteiskunnalle uusia mahdollisuuksia. Viime vuosina yhä useammin on saanut lukea palveluiden tehostamisesta, paikasta riippumattomista töistä, etäterveydenhuollosta ja tekoälyn potentiaalista opetuksessa, töiden tukena ja luovankin työn apuna. Toimintatapojen muutokset vaikuttavat käsi-

308 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 170.

309 Ks. esimerkiksi Mazur, O. (2018). Taxing the robots. *Pepp. L. Rev.*, *46*, 277.

310 Ks. esimerkiksi Van Parijs, P. (2004). Basic income: a simple and powerful idea for the twenty-first century. *Politics & Society*, *32*(1), 7–39.

311 Benito, A. (2006). Does job insecurity affect household consumption? *Oxford Economic Papers*, *58*(1), 157–181.

312 Ks. esimerkiksi Sosiaali- ja terveysministeriön lausunto hallituksen kaavailemaan toimeentulotuen rajoittamiseen liittyen: Sosiaali- ja terveysministeriö. (2023). Työmarkkinatuen ja toimeentulotuen uudistuksen vaikutukset. Hallitusneuvotteluihin annettu lausunto.

313 Hallitusohjelma 2023. Vahva ja välittävä Suomi.

tyksiin siitä, mikä on *tavanomaista*. Tavanomainen näyttäytyy usein yleisesti hyväksyttynä tai jopa järkevänä ja toivottuna[314]. Muutokset siinä, mitä yhteiskunnassa pidetään yleisesti hyväksyttynä tai toivottuna, vaikuttavat taas siihen, miten yksilöt ja instituutiot toimivat tai *voivat toimia* kohtaamatta paheksuntaa. Algoritmisaatio siis yhtäältä vaikuttaa käsityksiin siitä, mikä on järkevää, toisaalta muokkaa mahdollisuuksien horisonttia niin yksilöiden kuin yhteiskunnan tasolla. Kun käsitykset ovat muuttuneet ja uusia toimintatapoja otettu käyttöön, palaaminen takaisin aiempaan voi herättää suurta vastustusta[315] – vaikka uudet tavat paljastuisivat haitallisiksi. Tämän takia on perusteltua pyrkiä hillitsemään nopeita yhteiskunnallisia muutoksia: se antaa aikaa arvioida potentiaalisia vaikutuksia yhteiskuntien todellisuuteen ennen niiden realisoitumista, mikä luonnollisesti voi mahdollistaa joidenkin haittojen ennalta estämisen.

Poliittinen päätöksenteko ja demokraattiset prosessit ovat perinteisesti pyrkineet hidastamaan yhteiskunnan muutosta ja vakauttamaan yhteiskunnallisia prosesseja[316] ja siten myös hidastamaan käsitysten ja käytänteiden muutoksia. Yhteiskunnan muutostahdin kiihtyessä tämä ei kuitenkaan onnistu vaan eritahdistuminen demokraattisten sääntelykeinojen ja yhteiskunnallisten muutosten välillä voimistuu[317]: sen sijaan, että demokraattiset prosessit vakauttaisivat yhteiskuntaa ja mahdollistaisivat hallitun muutoksen, muutos tapahtuu monilla alueilla ennen demokraattista hyväksyntää ja valvontaa tai täysin demokraattisesta päätöksenteosta irrallaan[318]. Seurauksena demokratia heikkenee, millä voi olla laajoja vaikutuksia siihen, miten yhteiskunnat voivat toimia ja miten oikeutettuna yhteiskunnalliset käytänteet tai niiden uudistukset nähdään. Jos demokratia ei toteudu, on myös mahdollisesti helpompi viedä läpi muutoksia ja uudistuksia, joilla ei todellisuudessa ole kansan enemmistön tukea. Tällaiset muutokset voivat pahimmillaan laajasti heikentää kansalaisten hyvinvointia ja hyvän elämän edellytyksiä.

314 Politiikan tutkimuksessa puhutaan Overtonin ikkunasta: yleinen mielipide määrittää näkemyksiä siitä, mikä on mahdotonta – radikaalia – hyväksyttävää – järkevää – suosittua. Kun Overtonin ikkuna siirtyy tai sitä siirretään, käsitykset esimerkiksi hyväksyttävän rajoista muuttuvat. Teknologian ja kriisien vaikutuksesta Overtonin ikkunaan ks. esimerkiksi Doyle, D. (2020). The Overton Window has been flung wide open. (1.5.2020). *Medium.*

315 Bobric, G. D. (2021). The overton window: A tool for information warfare. *ICCWS 2021 16th International Conference on Cyber Warfare and Security* (p. 20). Academic Conferences Limited.

316 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

317 Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. Teoksessa *Temporal Boundaries of Law and Politics* (s. 72–87). Routledge.

318 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.

3.5.1.3 Vaikutukset luottamukseen

Mikäli yhteiskuntaan kohdistuvat haitat lisääntyvät ja ihmisten mahdollisuudet elää hyvää elämää heikentyvät laajasti niiden seurauksena, on oletettavaa, että ihmisten luottamus yhteiskuntaan heikkenee. Luottamus yhteiskuntaan ja etenkin sen poliittisiin ja oikeudellisiin instituutioihin on perusedellytys yleiselle luottamukselle (*generalized trust*)[319], jota tarvitaan, jotta yhteiskunta voisi toimia tehokkaasti ja hyvinvointia edistäen[320].

Monet toisistaan riippumattomat kehityskulut voivat heikentää ja mahdollisesti ovat jo heikentäneet ihmisten yleistä luottamusta. Esimerkiksi kuplautumista ja polarisaatiota voimistavat sosiaalisen median suosittelualgoritmit voivat lisätä yleistä epäluuloa niin instituutioita kuin muita ihmisiä kohtaan[321] samalla, kun algoritmit lisäävät disinformaation tuotannon ja levittämisen mahdollisuuksia[322]. Lisääntynyt disinformaatio voi heikentää luottamusta myös asiallisiin tietolähteisiin. Samalla algoritmisaatio antaa mahdollisuuksia tahalliseen mielipidevaikuttamiseen ja manipulointiin: on tutkittu, että esimerkiksi jotkut autoritääriset valtiot ovat pyrkineet vaikuttamaan muiden valtioiden vaalituloksiin ja poliittiseen asenneilmapiiriin sosiaalisen median bottien[323] ja ammattimaisten trollien[324] avulla[325]. Jotta demokratia voisi toteutua, ihmisillä on oltava ensinnäkin pääsy tiedon ääreen ja toisekseen halu ja

319 Rothstein, B., & Stolle, D. (2008). The state and social capital: An institutional theory of generalized trust. *Comparative politics*, *40*(4), 441–459.

320 Algan, Y. (2018). Trust and social capital. For Good Measure: Advancing Research on Well-Being Metrics Beyond GDP; Stiglitz, J., Fitoussi, J., Durand, M., Eds, 283–320.

321 Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298–320.

322 Zimmer, F., Scheibe, K., Stock, M., & Stock, W. G. (2019). Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice*, *7*(2), 40–53.

323 Botti viittaa tietokoneohjelmaan, joka toimii sille määriteltyjen toimintaohjeiden mukaan varsinkin suurta työmäärää, toistoa, jatkuvaa päivystystä ym. vaativissa tehtävissä. Tällaisia tehtäviä voivat olla esimerkiksi sosiaalisessa mediassa sisällön jakaminen, mainostaminen tai muu vastaava (sosiaaliset botit), tai toisaalta reaaliaikaisten internetkeskusteluiden käyminen, esimerkiksi nettisivujen asiakaspalvelussa (chatbotit). Ks. esimerkiksi Gehl, R. W., & Bakardjieva, M. (Eds.). (2016). *Socialbots and their friends: Digital media and the automation of sociality*. Taylor & Francis.

324 Trolli viittaa internetin keskusteluissa henkilöön, joka pyrkii ärsyttämään, aiheuttamaan ristiriitoja ja provosoimaan. Trollit jakavat usein rasistista, homo- ja transfoobista, misogyynista tai muuta vihaa lietsovaa sisältöä. Trollausta hyödynnetään poliittisen propagandan jakamiseen ja näkyvyyden lisäämiseen, ja se voi olla osa hybridivaikuttamista. Ks. esimerkiksi Hannan, J. (2018). Trolling ourselves to death? Social media and post-truth politics. *European Journal of Communication*, *33*(2), 214–226.

325 Ks. esimerkiksi Venäjän trollien vaikutuksesta USA:n presidentinvaaleissa 2016: Almond, D., Du, X., & Vogel, A. (2022). Reduced trolling on Russian holidays and daily US Presidential election odds. *Plos one, 17*(3), e0264507.

kyky tiedon hyödyntämiseen demokraattisissa prosesseissa[326]. Algoritmiset järjestelmät vaikuttavat näihin edellytyksiin. Seurannaishaittoja voi olettaa kohdistuvan myös yhteiskuntaan, jos demokratian edellytykset heikkenevät ja sen seurauksena vallanjaon oikeutus käy mahdollisesti kyseenalaiseksi[327].

Propagandan, valheiden ja harhaanjohtavan tiedon levittämisellä voi olla kauaskantoisia seurauksia. Suurin ongelma ei liene se, että valheita uskottaisiin laajassa mittakaavassa, vaikka disinformaation onkin osoitettu leviävän sosiaalisen median kaikukammioissa[328], mikä viittaisi siihen, että ainakin joissain ryhmissä siihen myös uskotaan. Lisääntyvä altistuminen valheelliselle tiedolle nakertaa kuitenkin myös pohjaa yhteiskuntajärjestystä ja oikeudenmukaisuutta ylläpitäviltä voimilta: tieteeltä, medialta ja demokratialta[329]. Lisäksi, kun käsitys tiedosta ja totuudesta käy epävarmaksi, ihmisen mahdollisuudet autonomiseen, tietoon perustuvaan päätöksentekoon heikkenevät ja epävarmuus ja epäluuloisuus lisääntyvät. Solidaarisuus ja usko omiin vaikutusmahdollisuuksiin vähenevät[330]. Tämä heikentää yhteiskunnan toimintamahdollisuuksia. Aihe on noussut niin akateemiseen kuin populaariinkin keskusteluun voimakkaasti viime vuosina, kun vastakkainasettelu on kiristynyt[331] ja demokratiaa, ihmisoikeuksia ja oikeusvaltiota kritisoivat äänenpainot ovat voimistuneet, mikä näkyy esimerkiksi kasvavana äärioikeiston ja fasististen arvojen kannatuksena[332].

Usko oikeudenmukaisuuden toteutumiseen ja luottamus yhteiskuntaan voivat olla koetuksella myös, mikäli yhteiskunnan päätöksenteossa automatisoitujen järjestelmien matemaattista logiikkaa noudattavat prosessit korvaavat intressipunninnan

326 Ibid.

327 Bayamlıoğlu, E., & Leenes, R. (2018). The 'rule of law' implications of data-driven decision-making: a techno-regulatory perspective. *Law, Innovation and Technology,* 10(2), 295–313;

Yeung, K. (2019). Responsibility and AI - A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe.

328 Zimmer, F., Scheibe, K., Stock, M., & Stock, W. G. (2019). Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice*, *7*(2), 40–53.

329 Arendt, H. (1973). The origins of totalitarianism. *New York*.

330 Coeckelbergh, M. (2022). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 1–10.

331 Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*;

Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118.

332 Mondon, A., & Winter, A. (2020). Reactionary democracy: How racism and the populist far right became mainstream. Verso Books.

ja inhimillisen harkinnan. Ihmisen ymmärrys päätöksen perusteista voi heikentyä ja mahdollisuudet vaikuttamiseen kaventua[333]. Tämä voi jälleen heikentää luottamusta yhteiskunnallisiin instituutioihin. Algoritmit poistavat päätöksenteosta moraalisen ulottuvuuden ja korvaavat pyrkimyksen tapauskohtaiseen harkintaan ja oikeudenmukaisuuteen pyrkimyksellä tehokkaaseen sääntöjen noudattamiseen ja optimoituun päätöksentekoon[334]. Yksilö häivytetään päätöksentekoprosessista ja pelkistetään arvioitavaksi ja optimoitavaksi dataksi[335]. Samaan aikaan myös oikeudellinen mielikuvitus väistämättä muuttuu; kun algoritmisten teknologioiden sääntelyssä keskitytään hallitsemaan teknologioiden suunnitteluprosessia, esimerkiksi oikeussuojan turvaaminen[336] ja oikeusvaltioperiaatteen toteuttaminen[337] käsitetään yhä laajemmin tulevaisuuteen suuntautuvina (teknologioiden) suunnittelukysymyksinä. Vaikutukset oikeusvaltioon ja oikeudenmukaisuuteen voivat olla perustavanlaatuisia, kuten myös Koivisto ja muut argumentoivat[338].

Yhteiskunnassa ja ihmisten elämissä huomaamattomasti ja läpitunkevasti vaikuttavien teknologioiden tehokas sääntely on välttämätöntä, sillä kuten Nemitz argumentoi, demokratia ei voi selvitä sääntelemättömien teknologioiden aikakaudella[339]. Teknologioiden aiheuttamat, yhteiskuntaan kohdistuvat seurannaishaitat voivat olla yllättäviä, kun eri vaikutukset kumuloituvat ja muokkaavat yhteiskunnan rakenteita monista suunnista ja monin eri tavoin. Mikäli elämä ja yhteiskunta muovataan mukailemaan teknologioita sen sijaan, että teknologioiden vaadittaisiin mukailevan demokraattisen yhteiskunnan periaatteita, valta yhteiskunnassa siirtyy yhä vahvemmin voittoa tavoittelevien yritysten käsiin. Seuraukset olisivat väistämättä perustavanlaatuisia. Uhkana on, että algoritmisten teknologioiden lisääntyvä, hallitsematon tuotanto ja käyttö johtavat yhteiskunnat kohti tulevaisuutta, jossa seuran-

333 Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

334 Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197–211.

335 Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. Konferenssijulkaisu, *Proceedings of the 2018 Chi conference on human factors in computing systems* (1–14).

336 Rommetveit, K., & Van Dijk, N. (2022). Privacy engineering and the techno-regulatory imaginary. *Social Studies of Science*, *52*(6), 853–877.

337 Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review*, *82*(3), 425–455.

338 Koivisto, I., Koulu, R., & Larsson, S. (2024). User accounts: How technological concepts permeate public law through the EU's AI Act. *Maastricht Journal of European and Comparative Law,* 1023263X241248469.

339 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. 10.

naishaitat rapauttavat yhteiskuntien toimintamahdollisuuksia tavoilla, joita emme pysty ennakoimaan emmekä hallitsemaan. Haitat voivat muuttua entistä vakavammiksi, mikäli sosioekonomiset muutokset heikentävät yhteiskunnan taloudellista tilannetta: samalla, kun algoritmisen transformaation hallitsemiseen ja potentiaalisten haittojen ja mahdollisten akuuttien kriisien hillitsemiseen tarvittaisiin enemmän resursseja, valtion varallisuus voi huveta automaation lisäämän työttömyyden vähentäessä verotuloja.

4 Tekoälyn ominaisuuksien vaikutukset haittojen syntymiseen

Edellisen luvun pohjalta on selvää, että tekoälyteknologioilla on potentiaalia aiheuttaa mitä moninaisempia haittoja. Tässä luvussa selvitän, miten teknologioiden tekniset ominaisuudet vaikuttavat haittapotentiaaliin. Sitä varten tarkastelen ensin erilaisia tapoja luokitella algoritmisia teknologioita ja selvitän niille tyypillisiä ominaisuuksia, minkä jälkeen arvioin, minkälaisia riskejä eri ominaisuuksiin kiinnittyy. Ensimmäisessä alaluvussa käyn tiiviisti lävitse, minkälaisia sovelluksia voidaan lukea ja on luettu tekoälyn piiriin, ja miten näitä on perusteltua jaotella analyysin helpottamiseksi. Toisessa alaluvussa keskityn nimenomaisesti tieto- ja logiikkapohjaisten järjestelmien ominaisuuksiin ja arvioin, minkälainen haittapotentiaali niihin liittyy. Neljännessä alaluvussa siirryn käsittelemään koneoppivien teknologioiden ominaisuuksia ja haittapotentiaalia.

4.1 Tekoälyn luokitteleminen

Kuten aiemmin on käynyt ilmi, tekoäly ei ole yksi yhtenäinen teknologia, vaan joukko mitä moninaisempia järjestelmiä, joita voidaan hyödyntää laajasti erilaisten tehtävien suorittamiseen. Näin ollen myös eri järjestelmien tekniset ominaisuudet ovat erilaisia. Sääntöihin perustuvat automaattiset päätöksentekojärjestelmät eroavat huomattavasti esimerkiksi online-alustojen suosittelutyökaluista tai päätöksenteon tukena käytettävistä koneoppivista ennustemalleista. Tekoälysovelluksia kuitenkin yhdistää kyky suorittaa tehtäviä, joiden suorittamiseen on yleensä tarvittu ihmistä.

Suomen tekoälystrategiassa tekoäly määritellään seuraavasti:

> Tekoäly tarkoittaa laitteita, ohjelmistoja ja järjestelmiä, jotka kykenevät oppimaan ja tekemään päätöksiä lähes samalla tavalla kuin ihmiset. Tekoälyn avulla koneet, laitteet, ohjelmat, järjestelmät ja palvelut voivat toimia tehtävän ja tilanteen mukaisesti järkevällä tavalla.[340]

[340] Työ- ja elinkenoministeriö. (2019). *Edelläkävijänä tekoälyaikaan – Tekoälyohjelman loppuraportti.* Työ- ja elinkeinoministeriön julkaisuja 2019:23. 14.

Tekoälystrategian määritelmässä tekoälyyn siis liitetään oppimiskyky. Usein puhutaan koneoppimisesta (*machine learning*). Koneoppivat tekoälyjärjestelmät on kehitetty koneoppimismenetelmillä ja ne pystyvät opetusdatasta oppimansa perusteella tunnistamaan esimerkiksi säännönmukaisuuksia, kaavoja ja/tai asioiden välisiä suhteita. Opetusdatasta oppimansa perusteella ne pystyvät käytössä ollessaan tekemään ennusteita tai arvioita niille syötetystä uudesta datasta. Käyttökontekstista riippuen koneoppivia algoritmeja voidaan hyödyntää esimerkiksi luottoriskin suuruuden arvioimisessa henkilöstä saatavilla olevien henkilötietojen perusteella, tai algoritmisten työkalujen avulla voidaan jopa tehdä lopullinen päätös siitä, onko luoton myöntämiselle edellytyksiä. Algoritmeja on liitetty mitä moninaisimpiin yhteyksiin: ne ehdottavat suoratoistopalveluissa käyttäjän profiiliin sopivia sarjoja, arvioivat terveysdatan perusteella riskejä eri sairauksille ja tunnistavat tai tuottavat kuvia, tekstiä tai ääntä.

Koneoppivien teknologioiden lisäksi tekoälyksi on usein perusteltua määritellä myös vähintään jokseenkin monimutkaiset sovellukset, jotka voidaan erottaa yksinkertaisemmista teknologisista järjestelmistä niiden adaptiivisuuden[341] tai niin sanotun itsenäisen toimintakyvyn[342] perusteella. Tällaisen määritelmän perusteella tekoälyksi voidaan katsoa paitsi järjestelmät ja sovellukset, jotka on kehitetty koneoppimismenetelmin, myös muilla menetelmillä kehitetyt sovellukset, joita voidaan käyttää ratkomaan verrattain monimutkaisia tehtäviä ja/tai jotka on ohjelmoitu reagoimaan ympäristössä tapahtuviin muutoksiin. Tällainen määritelmä ei rajaa tekoälyä puhtaasti sen teknisten ominaispiirteiden ja ohjelmoinnin tavan vaan sen tarjoamien käyttötapojen ja -mahdollisuuksien perusteella.

Tekoälyä voidaan luokitella myös sen perusteella, kuinka laaja-alaisesti se pystyy toimimaan. Heikko tai kapea-alainen tekoäly (*narrow AI*) suoriutuu tehtävistä tietyssä, ennalta määrätyssä kontekstissa. Kapea-alainen tekoälysovellus voi esimerkiksi suoriutua kuvien tai tekstin tulkitsemisesta, pelata shakkia tai mallintaa säätilan muutoksia, mutta se ei pysty toimimaan luovasti useilla eri tavoilla tai erilaisissa ympäristöissä. Käytännössä kaikki nykyinen tekoäly on määritelmällisesti kapea-alaista. Vahva tai yleistekoäly (*artificial general intelligence, AGI*), mikäli sellainen

341 Tällä viitataan algoritmisen teknologian kykyyn valita toimintansa sen perusteella, mitä se havaitsee ympäristössään. Kyseessä ei ole siis koneoppiminen vaan algoritmiin koodattu toimintaohje, jonka perusteella algoritmi reagoi määrätyillä tavoilla ympäristönsä muutoksiin. Esimerkiksi suosittelualgoritmit havainnoivat kulutustottumuksia ja suosittelevat samankaltaisia tuotteita/sisältöjä seuraavaksi.

342 Itsenäinen toimintakyky viittaa siihen, että järjestelmä pystyy toimimaan ilman ihmisen välitöntä ohjausta. Esimerkiksi Hildebrandt pitää itsenäistä toimijuutta (*agency*) tekoälyn tärkeänä ominaispiirteenä. Hildebrandt, M. (2015). *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing. 22–30.

onnistuttaisiin luomaan, toimisi ihmisen kaltaisesti ja pystyisi soveltamaan tietoa moninaisissa ympäristöissä ja tilanteissa. Mikäli yleistekoäly onnistutaan kehittämään, on arveltu, että se väistämättä johtaisi supertekoälyn (*artificial super intelligence*) kehittymiseen. Supertekoäly ylittäisi ihmisen kognitiivisen kapasiteetin kaikilla mahdollisilla tasoilla, minkä on spekuloitu pahimmillaan johtavan ihmiskunnan tuhoon[343]. Tällainen jaottelu on hyödyllinen etenkin tulevaisuuden skenaarioita käsiteltäessä, mutta nykyisten järjestelmien luokittelemisessa siitä ei ole juurikaan apua. Nykypäivänä käytössä olevat teknologiat lukeutuvat jaottelun heikon tekoälyn piiriin, eivätkä tutkijat ole yksimielisiä siitä, onko yleistekoälyn kehittäminen ylipäätään mahdollista.

Yleistekoäly on syytä erottaa *yleiskäyttöisestä* tekoälystä (*general purpose AI*). Yleiskäyttöinen tekoäly on nostettu erikseen esille esimerkiksi EU:n tekoälyasetuksen valmistelun aikana[344], ja sillä viitataan tekoälyjärjestelmiin, joita voi käyttää useissa eri konteksteissa. Tällaisia järjestelmiä ovat esimerkiksi suuret kielimallit ja kuvantunnistusjärjestelmät, jotka voidaan liittää moniin eri yhteyksiin. Yleistettävyydestään huolimatta tämän tyyppiset tekoälyjärjestelmät eivät määritelmällisesti ole yleistekoälyä; ne toimivat rajatulla alueella, ja sen ulkopuolella ne eivät pysty toimimaan luotettavasti tai ollenkaan. Esimerkiksi kuvantunnistukseen hyödynnettävä yleiskäyttöinen järjestelmä suoriutuu kuvien tunnistamisesta erilaisissa konteksteissa, mutta ei sen sijaan todennäköisesti pysty tuottamaan tekstiä tai pelaamaan shakkia. Yleiskäyttöiset järjestelmät ovat kuitenkin askel yleistekoälyn suuntaan: niitä on nimensä mukaisesti mahdollista käyttää monissa erilaisissa tehtävissä.

Jotta olisi mielekästä analysoida, miten erilaiset tekoälyjärjestelmien haitat kiinnittyvät tekoälyjärjestelmien ominaisuuksiin, tarvitaan jaottelua, joka mahdollistaa järjestelmien teknisten ominaisuuksien huomioimisen. EU:n alkuperäisessä tekoälysäädösehdotuksessa tekoälyksi määriteltiin logiikka- ja tietopohjaiset (*logic based, knowledge based*), koneoppivat (*machine learning*) teknologiat sekä tilastolli-

343 Jakoa on hyödynnetty etenkin filosofian tutkimuksen alueella; ks. supertekoälyn riskeistä esimerkiksi Bostrom, N. (2014). *Superintelligence.* Oxford University Press;
Kurzweil, R. (2022). Superintelligence and singularity. In Machine Learning and the City: Applications in Architecture and Urban Design, 579–601;
Russell, S. (2019). It's not too soon to be wary of AI: We need to act now to protect humanity from future superintelligent machines. *IEEE Spectrum*, *56*(10), 46–51;
Tegmark, M. (2018). Life 3.0: Being human in the age of artificial intelligence. Vintage.

344 Eurooppa-neuvosto. General approach on EU AI Act. (14954/22), art. 3(1b). Parlamentin kannassa puhutaan perustamalleista (*foundation model*), jotka rinnastuvat yleiskäyttöisiin tekoälysovelluksiin. Ks. Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Muutos 168. Art. 3(1c).

siin menetelmiin pohjaavat sovellukset[345]. Viimeisin on jätetty pois lopullisesta versiosta (resitaali 12). Vaikka jaottelu algoritmisen teknologian teknisen toteutustavan mukaan vaikuttaa jokseenkin kömpelöltä, kun tähtäimessä on arvioida teknologioiden riski- tai haittapotentiaalia, sen avulla pystyy kuitenkin luokittelemaan eri teknologioille tyypillisiä ominaisuuksia ja näiden ominaisuuksien vaikutuksia riskien ja haittojen muodostumisessa. Kiinnitän analyysini tekoälysäädöksen jaotteluun.

4.2 Tekoälyjärjestelmien ominaisuuksiin linkittyvä haittapotentiaali

Tekoälyyn liittyvien haittojen tutkimuksessa tekoälyn ominaisuuksien merkitys haittojen ilmenemiselle on vielä toistaiseksi jäänyt varsin vähälle huomiolle. Tekoälyyn liittyviä haittoja on tutkittu yhtäältä teoreettisella tai jopa spekulatiivisella tasolla supertekoälyn kontekstissa[346], toisaalta empiirisemmin esimerkiksi niin kutsuttujen mustien laatikoiden[347], vinoumien ja datan osalta[348]. Tekoälyjärjestelmien ominaisuuksien vaikutuksia järjestelmistä seuraaviin haittoihin on tutkittu vasta vähän. Viljanen[349] tarjoaa tähän kuitenkin hyvän lähtökohdan.

Viljanen on hahmotellut tekoälyjärjestelmien haittapotentiaaliin vaikuttavia seikkoja eli niitä tekoälyjärjestelmän ominaisuuksia, jotka nostavat riskiä tietyille, suoraan teknologiasta seuraaville haitoille, ja vaikeuttavat riskien hallitsemista. Näitä ovat hänen mukaansa järjestelmän itsenäisyys (*technological agency*), monimutkaisuus (*complexity*), tulkitsemattomuus tai selittämättömyys (*uninterpretability*), epälineaarinen toimintalogiikka (*non-linear system performance*), indeterminanssi (*indeterminacy*) eli mahdollisuus siihen, että samalla syötetyllä datalla saadaan erilaisia lopputulemia, sekä dynaamisuus (*dynamicity*) eli jatkuva käytössä oppiminen. Mitä useampi näistä ominaisuuksista on osa järjestelmää, sitä vaikeampi tekoälyjärjestelmän toimintaa on ennakoida ja kontrolloida. Tämä luonnollisesti lisää riskiä sille, että järjestelmät eivät toimi kuten niiden pitäisi, ja syntyy haittoja. Ominaisuuksien kumulaatio myös vaikeuttaa järjestelmien toimintavarmuuden tes-

345 Euroopan komissio. 2021/0106 (COD). Proposal for Artificial Intelligence Act. Annex I.

346 Ks. esimerkiksi Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

347 Ks. esimerkiksi Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

348 Ks. esimerkiksi O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

349 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

taamista, mikä todennäköisesti lisää haitallisten seurausten mahdollisuuksia ja niiden rajoittamisen vaikeutta.[350]

Koska tekoälyteknologioita käytetään monissa erilaisissa tarkoituksissa, myös niiden ominaisuuksien kirjo on vaihteleva. Erilaisten teknisten ratkaisujen riskien ja haittapotentiaalin tunnistaminen parantaa mahdollisuuksia paitsi vähentää haittoja, myös harkita ylipäätään, onko järjestelmän käytöstä seuraava haittojen riski kestävällä tasolla. On selvää, että mitä moninaisempia riskejä järjestelmän ominaisuuksista seuraa, sitä suurempi riski haitoille on. Haittojen vakavuuteen ja niiden muodostumisen todennäköisyyteen vaikuttavat kuitenkin ominaisuuksien lisäksi myös järjestelmän käyttöympäristö ja käytön tavat. Jos tekoälyn ominaisuuksista kumpuava riski haittojen ilmenemiselle arvioidaan korkeaksi, olisi perusteltua, että järjestelmää ei otettaisi käyttöön ainakaan ympäristössä, jossa riskin toteutuminen levittäisi vaikutukset hyvin laaja-alaisiksi ja/tai jossa riskin suuruusluokka olisi kestämätön. Käytännössä näin ei aina tapahdu joko siksi, että järjestelmiin tai niiden käyttöympäristöön kytkeytyvää haittapotentiaalia ei tunnisteta, tai siksi, että tunnistettua haittapotentiaalia ei syystä tai toisesta oteta riittävän vakavasti.

On epäselvää, kuinka vahvasti seurannaishaitat, esimerkiksi haitat oikeusvaltioperiaatteelle[351], demokratialle[352] ja oikeudenmukaisuudelle[353] kiinnittyvät algoritmisten järjestelmien teknisiin ominaisuuksiin. Todennäköisesti algoritmisen teknologian tekniset ominaisuudet ovat kuitenkin joissakin tilanteissa haitan syntymisen kannalta avainroolissa. Esimerkkinä mainittakoon ensimmäisessä osatutkimuksessa[354] esiin nostetut osakekaupassa hyödynnettävät algoritmit. Pörssikaupan algoritmien riski kiinnittyy hyvin selvästi niiden ominaisuuksiin, erityisesti itsenäiseen ja äärimmäisen nopeaan toimintaan, ja lisäksi pörssialgoritmien laumakäyttäytyminen lisää haittojen syntymisen potentiaalia. Pörssikaupan riskien realisoituminen voi pahimmillaan horjuttaa koko maailmantaloutta. Lisäksi on todennäköistä, että mahdollisesti seuraava taloudellinen taantuma vähentäisi esimerkiksi yhteiskunnallisia palveluita ja voisi altistaa yksilöt niin taloudellisille kuin terveydellisille haitoille, rajoittaa autonomiaa ja heikentää toiminnan mahdollisuuksia yhteiskunnassa. Haas-

350 Ibid.

351 Ks. esimerkiksi Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, *30*(3), 291–323;
Brownsword, R. (2016). Technological management and the rule of law. *Law, Innovation and Technology*, *8*(1), 100–140.

352 Ks. esimerkiksi Coeckelbergh, M. (2022). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 1–10.

353 Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

354 Malik, H. M., Viljanen, M., Lepinkainen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

teet algoritmisten järjestelmien toiminnan ennakoitavuudessa ja kontrolloinnissa, joita riskialttiit ominaisuudet lisäävät, vaikuttavat myös välittömiin haittoihin kiinnittymättömien seurannaishaittojen muodostumiseen. Epäsuorien seurannaishaittojen syntyminen ei riipu niinkään tekoälyjärjestelmien virheellisestä tai yllättävästä toiminnasta vaan järjestelmän käytön tavoista ja käyttökontekstista. Näin ollen on todennäköistä, että tekoälyteknologioiden tekniset ominaisuudet eivät määritä seurannaishaittoja yhtä vahvasti kuin välittömiä haittoja.

Keskityn seuraavaksi arvioimaan, miten erilaisille tekoälyteknologioille tyypilliset ominaisuudet vaikuttavat erilaisten yhteiskunnallisten haittojen muodostumiseen. Teknologiat jaotellaan EU:n tekoälyasetuksen pohjalta tieto- ja logiikkapohjaisiin järjestelmiin sekä koneoppiviin järjestelmiin. Näihin jokseenkin karkeisiin kategorioihin kuuluu monenlaisia teknologioita, joissa erilaiset ominaisuudet ja niihin kiinnittyvät riskit korostuvat. Jotta ominaisuuksien jaottelu ja arvioiminen kategorioiden perusteella olisi mielekästä, tulee kategorioille tyypilliset piirteet ensin pyrkiä ymmärtämään. Tämän takia arvioin molempien kategorioiden kohdalla ensin, minkälaisia tekoälyteknologioita EU:n tekoälyasetuksen pohjalta niihin vaikuttaisi sisältyvän.

4.3 Tieto- ja logiikkapohjaiset järjestelmät ja niiden haittapotentiaali

Yleisellä tasolla voidaan sanoa, että tieto- ja logiikkapohjainen tekoäly pohjautuu ennalta koodattuihin sääntöihin ja/tai tietokoneluettaviksi symboleiksi muunnettuun datapohjaan, jonka pohjalta järjestelmät käsittelevät tietoa ja ratkovat ongelmia. Jotta tämä olisi mahdollista, tieto täytyy muuntaa ensin koneluettavaan muotoon. Järjestelmä ohjelmoidaan etsimään tiedosta haluttuja seikkoja, minkä perusteella se pystyy muodostamaan ratkaisuja. Tällaisia järjestelmiä ovat esimerkiksi jotkut automaattisen päätöksenteon sovelluksista. Tieto- ja logiikkapohjaisia sovelluksia käytetään myös muun muassa suunnittelun, tiedonhaun ja -järjestelyn työkaluina sekä päättelyn apuvälineinä.

Suomessa hallintolaki mahdollistaa automaattisen päätöksenteon julkishallinnossa juuri tieto- ja logiikkapohjaisten järjestelmien osalta: hallintolain 53a § mukaan hallinnollisessa päätöksenteossa voidaan hyödyntää järjestelmiä, jotka ennalta koodattujen sääntöjen pohjalta konepohjaisen päättelyn keinoin tekevät ratkaisuja.

EU:n tekoälysääntelyn kehitystyön aikana tieto- ja logiikkapohjaisia järjestelmiä on määritelty useilla tavoilla. Neuvosto esitti tieto- ja logiikkapohjaisille järjestelmille seuraavaa määritelmää:

> 6(b) Logic- and knowledge based approaches focus on the development of systems with logical reasoning capabilities on knowledge to solve an application

> problem. Such systems typically involve a knowledge base and an inference engine that generates outputs by reasoning on the knowledge base. The knowledge base, which is usually encoded by human experts, represents entities and logical relationships relevant for the application problem through formalisms based on rules, ontologies, or knowledge graphs. The inference engine acts on the knowledge base and extracts new information through operations such as sorting, searching, matching or chaining. Logic- and knowledge based approaches include for instance knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning, expert systems and search and optimisation methods. [355]

Määritelmä oli melkoisen epäselvä, etenkin, kun sääntelyehdotuksen resitaalissa 6 rajattiin ehdotetun sääntelyn piiristä pois puhtaasti luonnollisten henkilöiden määrittelemiin sääntöihin pohjautuvat järjestelmät[356]. Resitaalissa 6(b) taas mukaan otettiin tieto- ja logiikkapohjaiset järjestelmät, jotka toimivat järjestelmään syötetyn tietomassan ja ennalta määriteltyjen sääntöjen pohjalta päättelymoottorin avulla. Ensi silmäyksellä näiden välillä vaikuttaisi olevan ristiriita. Todennäköisesti kuitenkin tällä pyrittiin tuomaan sääntelyn piiriin järjestelmät, joissa päättelymoottorin hyödyntämä sääntöpohja on rakennettu *muutoin kuin* luonnollisten henkilöiden työnä. Epäselväksi jäi, mitä tämä tarkoittaisi käytännössä. Kenties tämän epäselvyyden takia lopullisen säädösversion resitaalissa 12 tekoälyjärjestelmät päätettiin määritellä yksinkertaisemmin:

> A key characteristic of AI systems is their capability to infer. This capability to infer refers to the process of obtaining the outputs, such as predictions, content, recommendations, or decisions, which can influence physical and virtual environments, and to a capability of AI systems to derive models or algorithms, or both, from inputs or data. The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved. The capacity of an AI system to infer transcends basic data processing by enabling learning, reasoning or modelling.

355 Eurooppa-neuvosto. General approach on EU AI Act. (14954/22). Myös parlamentti katsoi kannassaan, että tekoälysääntelyn piiriin kuuluvat myös tietopohjaiset järjestelmät, etenkin erilaisissa hybridisysteemeissä.

356 Eurooppa-neuvosto. General approach on EU AI Act. (14954/22). Resitaali 6: "A system that uses rules defined solely by natural persons to automatically execute operations should not be considered an AI system."

Erotukseksi yksinkertaisista sovelluksista tekoälyn ulkopuolelle rajattiin järjestelmät, jotka perustuvat pelkästään ihmisen määrittelemiin päättelysääntöihin (resitaali 12).

Lähtökohtaisesti tieto- ja logiikkapohjaiset järjestelmät toimivat if/then-logiikan mukaisilla päättelysäännöillä. Päättely on siis ainakin teorian tasolla mahdollista purkaa ja kuvata muodossa, jossa ihminen voi sen ymmärtää. Järjestelmän monimutkaisuuden kasvaessa se voi kuitenkin käydä vaikeaksi. On myös mahdollista, että monimutkaiset päättelysäännöt luodaan valtavien datamassojen pohjalta esimerkiksi juuri tekoälyjärjestelmien avulla, jolloin järjestelmän päättelysääntöjen logiikkaa voi olla äärimmäisen vaikeaa ymmärtää. Siitä huolimatta tieto- ja logiikkapohjaiset järjestelmät toimivat melko epätodennäköisesti tavoilla, joita ei olisi ollenkaan mahdollista ennakoida.

Viljasen[357] haitoille altistavien piirteiden joukosta tieto- ja logiikkapohjaisten tekoälyjärjestelmien mahdollinen *itsenäinen toiminta* on haitoille altistava ja niiden ennaltaehkäisemistä vaikeuttava tekijä. Itsenäisyys viittaa siihen, että järjestelmä saa vaikutteita ja pystyy hyödyntämään niitä toiminnassaan ilman ihmisen myötävaikutusta. Siitä huolimatta itsenäiset järjestelmät tekevät juuri sitä, mitä ne on ohjelmoitu tekemään. Itsenäisyys merkitsee siis sitä, että järjestelmä pystyy suoriutumaan ohjelmointinsa mukaisesta toiminnasta ilman ihmisen välitöntä osallistumista. Itsenäisyys tekee tekoälyjärjestelmistä houkuttelevia ja äärimmäisen hyödyllisiä työvälineitä monissa eri konteksteissa. Kun järjestelmä toimii itsenäisesti, nopeasti ja väsymättä, toimintoja on mahdollista tehostaa.

Inhimillisen osallistumisen vaikutus algoritmisten järjestelmien riskipotentiaalin hallinnassa on paljon tutkittu aihepiiri[358]. Jos algoritminen järjestelmä vapautetaan toimimaan ilman inhimillistä valvontaa[359], luonnollisesti myös mahdollisuudet tunnistaa ja keskeyttää haitallinen tai virheellinen toiminta heikkenevät. Itsenäisesti toimivien järjestelmien toiminnasta puuttuvat ihmisvetoisille järjestelmille tyypilliset turvallisuutta lisäävät seikat, kuten päätösten hajauttaminen useille tekijöille ja päätösperusteiden jatkuva uudelleen arvioiminen. Lisäksi tekoälyvetoiset järjestelmät voivat toimia niin nopeasti, että ihmisen mahdollisuudet valvoa toimintaa jäävät hei-

357 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

358 Ks. esimerkiksi Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153.

359 Jotkut argumentoivat, että ihminen tulisi pitää mukana järjestelmän toiminnassa (human-in-the-loop), toiset katsovat, että riittää, että ihmisellä on viimesijainen kontrollointimahdollisuus (human-on-the-loop). Järjestelmän itsenäisen toiminnan mahdollistava vaihtoehto on, että ihminen poistetaan järjestelmän toiminnasta kokonaan (human-out-of-the-loop).

koiksi, vaikka teoriassa ihmisellä olisikin mahdollisuus vaikuttaa järjestelmän toimintaan. Vaikka ihminen huomaisi järjestelmän virheellisen toimintalogiikan sinänsä nopeasti, on mahdollista, että vaikutukset ovat jo levinneet äärimmäisen laajalle alueelle. Tehokas puuttuminen mahdollisesti ilmeneviin ongelmiin käy siten hyvin vaikeaksi[360].

Kuten Malikin kanssa toisessa osatutkimuksessa esitimme, Suomessa automaattisen päätöksenteon mahdollistavan lainsäädäntöhankkeen aikana argumentoitiin vahvasti tekoälyn käytön puolesta, vaikka lopullisessa sääntelyssä sallittiinkin ainoastaan sääntöpohjainen automaatio. Laajempaa tekoälyn hyödyntämisen mahdollistamista kannatettiin esimerkiksi sillä perusteella, että konepohjainen järjestelmä tuottaa yhtenäisiä ratkaisuja, noudattaa systemaattisesti koodiaan ja siten estää inhimillisten virheiden syntymisen[361]. Sinänsä argumentaatio on perusteltua. Tekoälyn avulla toimivat järjestelmät lähtökohtaisesti poistavat erilaiset tulkinnat ja tulkintakäytännön heilahtelut ja soveltavat jokaiseen tapaukseen systemaattisesti samoja sääntöjä. Systemaattisuus lisää lopputulosten johdonmukaisuutta, mikä voi olla monissa tilanteissa tavoiteltavaa. Toisaalta se väistämättä myös vähentää päätöksenteon joustavuutta. Systemaattisuus myös luo riskin sille, että virheellinen toimintalogiikka laajenee koskettamaan kaikkia niitä, joihin tekoälyjärjestelmä vaikuttaa: kun tekoälyvetoinen järjestelmä toimii aina täysin samoilla perusteilla, virheelliset päättelysäännöt johtavat siihen, että jokainen syntyvä päätös, jossa virheellisiä päättelysääntöjä sovelletaan, on väistämättä perusteiltaan virheellinen[362].

Kun automaattisten tekoälyjärjestelmien itsenäisyyttä kasvatetaan, niiden toimintanopeutta voidaan lisätä. Tällä on mahdollisia vaikutuksia paitsi järjestelmän toimintaan, myös niihin prosesseihin, joissa järjestelmää käytetään. Mikäli itsenäisiä tekoälyjärjestelmiä otetaan käyttöön päätöksenteossa, päätösten ajalliset parametrit väistämättä muuttuvat[363]. Päätöksenteon nopeus voi lisääntyä ainakin kolmesta syystä. Ensinnäkin algoritmiset järjestelmät pystyvät etsimään ja keräämään tietoa ja tunnistamaan nopeammin tilanteita, joissa päätöksiä tarvitaan. Tällöin viive päätöstarpeen ilmenemisen ja päätöksenteon välillä lyhenee. Toisekseen algoritmit pystyvät yhdistelemään tietoa ja soveltamaan sääntöjä nopeasti, jolloin itse päätöksen-

360 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

361 Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

362 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

363 Susser, D. (2022, June). Decision Time: Normative Dimensions of Algorithmic Speed. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1410–1420).

teko nopeutuu. Kolmanneksi, jos prosessi toimii ilman ihmisen myötävaikutusta, tieto päätöksestä siirtyy sekä asianosaisille että mahdollisesti eri rekistereihin välittömästi. Kun nopeus kaikissa kolmessa vaiheessa lisääntyy samalla, kun järjestelmien itsenäinen toiminta poistaa turvallisuutta lisääviä varmuuskerroksia, kuten hajautuksen ja jatkuvan uudelleenarvioinnin, haittoja voi syntyä monissa eri vaiheissa tapahtuvien mahdollisten virheiden seurauksena tai eri järjestelmien yllättävinä yhteisvaikutuksina. Haitat voivat levitä nopeasti, ja haitalliseen toimintalogiikkaan puuttuminen voi viivästyä huomattavasti.

Algoritmisen järjestelmän itsenäisyys siis lisää riskiä etenkin automaattisten päätöksentekojärjestelmien mahdollisten virheiden systematisoitumiselle ja nopealle leviämiselle[364] ja sitä kautta virheellisten päätösten aiheuttamille välittömille haitoille. Kuten edellisessä luvussa argumentoitiin, vinoutunut tai virheellinen päätöksentekologiikka altistaa yksilöt ja ryhmät oikeudenloukkauksille, syrjinnälle ja työläille muutoksenhakuprosesseille, jotka voivat eri konteksteissa aiheuttaa erilaisia välittömiä haittoja. Välittömien haittojen seurauksena voi syntyä erilaisia seurannaishaittoja, mikäli välittömät haitat esimerkiksi heikentävät ihmisten yleistä luottamusta, lisäävät rakenteellista syrjintää tai kiristävät eri ryhmien asenteita tai arvoja. Vaikka algoritmisten järjestelmien suoriutumista tarkkaillaan ja valvotaan – kuten EU:ssa on uuden lainsäädännön myötä tehtävä – itsenäisesti toimivan järjestelmän riskien realisoitumista on vaikea estää täydellisesti: mahdollisuudet inhimilliseen valvontaan ovat väistämättä lähinnä pistokokeenomaisia, sillä säännönmukainen valvonta poistaa tekoälyn nopeuden ja tehokkuuden mukanaan tuoman hyödyn[365].

Koska säännönmukainen jälkikäteinen valvonta on valtavan nopeasti toimivien järjestelmien tapauksessa parhaimmillaankin vaillinaista, haittapotentiaalin hillitsemiseksi on usein ehdotettu laadun varmistamista sääntelemällä tekoälyjärjestelmien tuotantoprosessia niin sanotulla metasääntelyllä[366]. Tällaisella sääntelyllä luodaan tuotantoprosessille vaatimukset, joilla pyritään varmistamaan, että järjestelmä on suunniteltu ja toteutettu laadukkaasti ja sen toimintalogiikan asianmukaisuus varmistettu riittävällä tavalla ennakkoon. Lisäksi voidaan määritellä myös jälkikäteisen valvonnan kriteerit. Itse järjestelmän ominaisuuksiin sääntelyllä ei kuitenkaan puu-

364 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1). 192.

365 Suuret teknologia-alan yritykset ovat vedonneet samaan. Tällä ne ovat pyrkineet oikeuttamaan heikon algoritmisten järjestelmien valvonnan ja argumentoimaan tiukempia velvoitteita vastaan. Ks. esimerkiksi Yeung, K. (2019). Responsibility and AI - A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe. 39.

366 Viljanen, M. (2017). Algoritmien haaste: uuteen aineelliseen oikeuteen? *Lakimies 115 (2017): 7–8,* 1070–1087.

tuttaisi. EU:n tekoälysäädös on suurelta osin juuri metasääntelyn varaan rakentunut kokonaisuus.

Monien tieto- ja logiikkapohjaisten järjestelmien kohdalla metasääntely voikin olla riittävää, sillä ennakoitavuutta ja kontrolloitavuutta haastavia ominaisuuksia tällaisissa järjestelmissä on rajallisesti. Toisaalta, kun järjestelmien *monimutkaisuus* lisääntyy, niin tuotantoprosessinaikainen haittapotentiaalin tunnistaminen kuin riskien ennakoiminen käyvät yhä vaikeammiksi. Tieto- ja logiikkapohjaistenkin järjestelmien toimintaa voi ohjata valtava määrä koodia, jota on usein yhdistelty useista eri lähteistä[367]. Voi olla, että kokonaisuutta ei hahmota kukaan, jolloin järjestelmän toiminnan yksityiskohtainen ymmärtäminen ja hallitseminen voi olla äärimmäisen vaikeaa.

Laajat ja monimutkaiset tietokannat ja päättelysäännöt järjestelmien toiminnan taustalla nostavat riskitasoa. On mahdollista, että päättelysäännöt tai tietokannat, joita tieto- tai logiikkapohjainen järjestelmä käyttää, on luotu monimutkaisten ja mahdollisesti läpinäkymättömien ja selittämättömien koneoppivien ennustemallien *avulla*, ja/tai että järjestelmän päättely on niin sanotusti sumeaa (*fuzzy logic*)[368]. Tällöin myös sääntöpohjaisten järjestelmien toiminta voi noudattaa niin monimutkaista logiikkaa, että riskien hahmottaminen käy ihmiskognitiolle mahdottomaksi. Tämä lisää edelleen riskiä virheelliselle toiminnalle ja siitä mahdollisesti seuraaville välittömille ja seurannaishaitoille. Käyn monimutkaisuuden ja selittämättömyyden mukanaan tuomia riskejä laajemmin lävitse seuraavaksi, kun käsittelen koneoppivien järjestelmien haittapotentiaalia.

4.4 Koneoppivat järjestelmät ja niiden haittapotentiaali

Koneoppivat järjestelmät tarkoittavat järjestelmiä, joiden luomiseksi on hyödynnetty koneoppimismenetelmiä. Tällaisten järjestelmien luominen vaatii lähtökohtaisesti valtavat määrät käyttökelpoista dataa. Eri menetelmien avulla järjestelmät oppivat tunnistamaan opetusdatasta esimerkiksi säännönmukaisuuksia, malleja tai korrelaatioita ilman, että tunnistukseen vaadittavia seikkoja nimenomaisesti ohjelmoidaan algoritmiin. Opittuaan ne pystyvät yleistämään oppimansa ja sen perusteella tekemään ennusteita niille syötetystä uudesta datasta.

367 Puhutaan monien käsien ongelmasta (*problem of many hands*). Ks. esimerkiksi Coeckelbergh, M. (2019). Artificial intelligence: some ethical issues and regulatory challenges. *Technology and regulation*, *2019*, 31–34.

368 Tällöin totuusarvo voi olla mikä vain luku nollan ja yhden väliltä sen sijaan, että se olisi joko 1 (tosi) tai 0 (epätosi). Eli sen sijaan, että ohjelmointi noudattelisi kaavaa ”if X=1 then…” se noudattelisikin esimerkiksi kaavaa ”if X<0,6 then…”.

Koneoppimispohjaisissa järjestelmissä on valtavasti vaihtelua, samoin kuin opetuksen toteuttamisen tavoissa. Järjestelmien tuotannossa voidaan hyödyntää esimerkiksi ohjattua (*supervised*), ohjaamatonta (*unsupervised*) ja vahvistusoppimista (*reinforcement learning*)[369]. Erilaiset menetelmät mahdollistavat koneoppimiseen pohjautuvien järjestelmien moninaisuuden sekä laajat ja vaihtelevat käyttötavat.

Koneoppimispohjaisten järjestelmien teknisissä ominaisuuksissa on myös paljon vaihtelua, mutta on teoriassa mahdollista, vaikka käytännössä jokseenkin epätodennäköistä, että yksittäinenkin tekoälyjärjestelmä on itsenäisesti toimiva, monimutkainen, vaikeasti tulkittavissa ja selitettävissä, epälineaarisesti toimiva, indeterministinen sekä dynaaminen, jolloin oppiminen tapahtuu myös käytön aikana. Käytännössä siis koneoppiviin järjestelmiin kytkeytyy niiden mahdollisista teknisistä ominaisuuksista kumpuava huomattava haittapotentiaali[370]: mitä enemmän riskiä lisääviä ominaisuuksia, sitä vaikeampi on ymmärtää järjestelmän toimintalogiikkaa ja varmistua sen asianmukaisuudesta esimerkiksi erilaisin testein tai simulaatioin, mikä luonnollisesti vaikeuttaa järjestelmän toiminnan ennakoimista ja kontrollointia, lisää riskiä virheelliselle toiminnalle ja siitä seuraaville haitoille sekä hidastaa haittoihin puuttumista.

Itsenäisyyden osalta pätevät jokseenkin samat seikat kuin mitä tieto- ja logiikkapohjaisten järjestelmien kontekstissa nostettiin esille: itsenäistä järjestelmää on vaikeampi kontrolloida ja sen toiminnassa tapahtuvia virheitä voi olla mahdotonta huomata riittävän nopeasti. Lisäksi turvallisuutta lisääviä, ihmisvetoisille järjestelmille luonteenomaisia mekanismeja on vähemmän. Tällaisia ovat esimerkiksi tehtävien hajauttaminen ja käytäntöjen jatkuva uudelleenarviointi. On tosin selvää, että haittojen riski kasvaa sitä mukaa, mitä enemmän ilman ihmisen myötävaikutusta toimivissa järjestelmissä on muitakin riskitasoa nostavia ominaisuuksia.

Koneoppivat järjestelmät ovat usein huomattavan monimutkaisia. Etenkin syviin neuroverkkoihin pohjaavat järjestelmät ovat laajoilla datamassoilla opetettuja ja äärimmäisen monimutkaisia järjestelmiä. Neuroverkkojen on tarkoitus imitoida ihmisaivojen toimintaa: niissä on syötekerroksen (input layer) ja ulostulokerroksen (output layer) lisäksi vaihteleva määrä piilokerroksia (hidden layer), joissa päättely tapahtuu. Kerrokset rakentuvat neuroneista (neuron), joihin kuhunkin liittyy muutettavia parametrejä (parameter), jotka määrittävät neuronien väliset painoarvot (weight) ja neuronikohtaisen kynnysarvon (bias). Syvissä neuroverkoissa kerroksia voi olla valtava määrä, ja kussakin kerroksessa lukemattomia neuroneita. Kerros ker-

369 Ks. tavoista lisää Ayodele, T. O. (2010). Types of machine learning algorithms. Teoksessa *New advances in machine learning*. Zhang, Y. (Ed.). 19–48.

370 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

rokselta neuronien määrä kasvaa, ja lopulta niitä on niin paljon, että ihmiselle mahdotonta ymmärtää niiden välisiä kytköksiä ja painoarvoja. Tämä lisää riskiä virheille, vinoumille ja epätarkoituksenmukaisille lopputulemille, ja vaikeuttaa järjestelmien testaamista sekä toiminnan asianmukaisuuden arviointia[371]. Luonnollisesti vaikeasti ymmärrettävän järjestelmän toiminnasta on myös vaikeaa havaita ja tunnistaa virheitä. Tämän seurauksena virheiden ja virheellisestä toiminnasta seuraavien välittömien haittojen riski kasvaa. Tämä on jokseenkin varmaa etenkin, jos tehokkaita tapoja reagoida mahdollisiin virheisiin ei ole suunniteltu osaksi järjestelmän käyttöä.

Tulkitsemattomuus kasvattaa järjestelmien riskitasoa entisestään. Esimerkiksi syviin neuroverkkoihin perustuvien järjestelmien piilokerroksissa tapahtuvaa päättelyä ei useinkaan ole mahdollista tulkita tai selittää tavalla, joka olisi ihmiselle ymmärrettävissä. On mahdollista, että niiden toimintaa voidaan arvioida lähinnä syötteen ja lopputuloksen perusteella, jolloin puhutaan mustista laatikoista[372]. Mustien laatikoiden potentiaalisesti ongelmallisen toiminnan tunnistaminen, paikantaminen ja todistaminen on vaikeaa, mikä pahimmillaan vaikeuttaa tai jopa estää haittoja aiheuttavien mekanismien muuttamisen. Etenkin tilanteissa, joissa järjestelmät myös toimivat näkymättömissä[373] eli niiden olemassaoloa on ylipäätään vaikea havaita, kokonaiskuva voi hämärtyä siten, että ongelmiin puuttuminen käy mahdottomaksi[374].

Mikäli tekoälyjärjestelmän suorittaman päättelyn perusteita ei ole mahdollista esittää ihmiselle ymmärrettävässä muodossa, on myös väistämättä vaikeaa varmistua siitä, että järjestelmä toimii oikeudenmukaisesti ja lakien ja säädösten puitteissa[375]. Selittämättömien järjestelmän toimintaa voi kuitenkin pyrkiä arvioimaan eri tavoin testaamalla, esimerkiksi simulaatioiden avulla. Laadukas testaaminen voi auttaa var-

371 Englanniksi puhutaan testaamisesta, arvioimisesta, validoinnista ja verifioinnista (testing, evaluation, validation and verification). Viittaan testaamisella ja toiminnan asianmukaisuuden varmistamisella prosessiin, joka pitää sisällään nämä vaiheet. Ks. esimerkiksi autonomisten asejärjestelmien testaamiseen liittyen Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. Teoksessa *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm,* 175–191. Cham: Springer International Publishing.

372 Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

373 Elliott, A. (2019). The culture of AI: Everyday life and the digital revolution. Routledge. xxi.

374 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

375 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

mistamaan, että toiminta on asianmukaista, tunnistamaan riskejä ja lisäämään mahdollisuuksia ennakoida ja vähentää mahdollisia haittoja[376].

Vaikka testaaminen olisi asianmukaista, tulkitsemattomien ja läpinäkymättömien järjestelmien liittäminen konteksteihin, joissa järjestelmän toiminnan perusteiden ymmärtäminen on tärkeää, johtaa herkästi epäluuloon ja luottamuksen rapistumiseen. Näin voi tapahtua etenkin, jos järjestelmän toiminta johtaa epäoikeudenmukaisiin lopputuloksiin, mutta mahdollisesti myös siksi, että järjestelmän tulkitsemattomuus voi *antaa syyn epäillä*, ettei järjestelmän toimintalogiikka kestä päivänvalossa tarkastelua. Lähtökohtaisesti päätösten hyväksyttävyyttä arvioitaessa perustelut ovat ihmisille usein erittäin olennaisia: ne auttavat arvioimaan päätösten oikeudenmukaisuutta ja sitä kautta rakentavat niiden legitimiteettiä. Tekoälyjärjestelmien tekemien päätösten perusteita ei kuitenkaan aina pystytä kuvaamaan ihmiselle ymmärrettävässä muodossa. Tämä voi osaltaan lisätä riskiä paitsi toimintalogiikan mahdollisesti tunnistamatta jääneistä virheistä kumpuaville välittömille haitoille, myös mahdollisesti lisääntyvästä luottamuspulasta seuraaville, yhteiskuntaan ja sen instituutioihin kohdistuville seurannaishaitoille. Seurannaishaittojen riski kasvaa, mikäli selittämättömiä järjestelmiä liitetään yhteiskunnallisesti herkkiin ympäristöihin, joissa niiden toiminta asettaa yhteiskunnallisten toimintojen oikeudenmukaisuuden kyseenalaiseksi. Lisäksi, mikäli algoritmisen järjestelmän toiminnan tulkitseminen ja selittäminen käyvät mahdottomiksi, järjestelmän mahdollisesti aiheuttamia haitallisia vaikutuksia voi olla huomattavan vaikea ennakoida. Tulkitsemattomuus luo myös tilaa algoritmisten järjestelmien aikaansaamien haittojen ja epäoikeudenmukaisen toiminnan opportunistiselle tai tahattomalle sivuuttamiselle.

Epälineaarisuudella viitataan tässä yhteydessä siihen, algoritmisen järjestelmän lopputulos ei muutu samassa suhteessa syötteen kanssa: pieni muutos syötteessä voi aiheuttaa suuren muutoksen lopputuloksessa tai päin vastoin. Suuri osa koneoppivista järjestelmistä toimii epälineaarisesti: esimerkiksi kuvantunnistusalgoritmien (*image recognition*) toiminnassa epälineaarisuus parantaa niiden suoriutumista yksinkertaisesti sen takia, että maailmassa monikaan asia ei noudattele lineaarista logiikkaa[377]. Epälineaarisuus tekee kuitenkin järjestelmän toiminnasta vaikeammin ennakoitavaa, mikä vaikeuttaa toimintalogiikan ymmärtämistä ja hallintaa. Näin ollen epälineaarisuus kasvattaa jälleen riskejä, joita tekoälyjärjestelmiin liittyy: se vaikeuttaa toimintalogiikan asianmukaisuuden varmistamista, virheiden tunnistamista ja järjestelmän toiminnan hallintaa.

Riittävän laajan testaamisen merkitys korostuu epälineaaristen järjestelmien kohdalla, sillä testaamalla voidaan selvittää esimerkiksi, mitkä raja-arvot syötteessä

376 Ibid.

377 Hijazi, S., Kumar, R., & Rowen, C. (2015). Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*, *9*(1).

ovat lopputuloksen kannalta määrittäviä, tai mikäli tämä ei ole mahdollista, testaamisen volyymiä voidaan kasvattaa siihen asti, että yllätyksiä ei enää ilmene. Järjestelmien laadukas testaaminen ja toiminnan asianmukaisuuden varmistaminen voivat vaatia esimerkiksi laajoja simulaatioita, joissa erilaisia skenaarioita voidaan arvioida riittävän monipuolisesti[378]. Tämä vaatii huomattavasti resursseja[379]. Vaikka testaaminen olisi laaja-alaista ja monipuolista, on mahdollista, ettei siinä osata silti ennakoida kaikkia relevantteja ympäristötekijöitä tai ympäristön muutoksia. Luonnollisesti mahdolliset puutteet ennakoinnissa lisäävät riskiä järjestelmän virheelliselle toiminnalle, ja riski virheiden aiheuttamille haitoille kasvaa. Mahdollisten haittojen ilmenemismuodot riippubat sekä teknologiasta että sen käyttökontekstista; esimerkiksi kuvan- tai hahmontunnistusalgoritmeja (*image/object recognition*) käytetään monissa eri yhteyksissä, ja autonomisen ajoneuvon virheellinen havainnointi voi johtaa hyvin erilaisiin haittoihin kuin esimerkiksi sosiaalisen median alustojen sisällönvalvontaan käytettävien algoritmien virheet.

Valtavillakaan resursseilla ei ole välttämättä mahdollista selvittää kattavasti, miten indeterminantit järjestelmät toimivat. Tämä tarkoittaa sitä, että virheellisen tai odottamattoman toiminnan ennakoiminen ja ennalta estäminen käy jälleen vaikeammaksi. Indeterminanttien järjestelmien riittävä testaaminen riskien tunnistamiseksi ja hallitsemiseksi voi olla mahdotonta, ja järjestelmien kontrolloiminen vaikeaa[380], sillä järjestelmät voivat päätyä aina uusiin lopputuloksiin ilman, että syötettä muutetaan. Aina vain uudet lopputulokset perustuvat järjestelmiin sisäänrakennettuun satunnaisuuteen. Esimerkiksi suuret kielimallit ovat lähtökohtaisesti indeterministisiä. Ne on opetettu valtavilla datamäärillä, ja niiden toimintaa ohjaavat syvät neuroverkot, minkä pohjalta ne generoivat sanoja toistensa perään tasapainotellen tarkoituksenmukaisuuden ja satunnaisuuden välillä. Satunnaistuksen ansiosta ne pystyvät tuottamaan loputtomasti erilaisia lopputulemia. Syvät neuroverkot ovat, kuten aiemmin esitetty, inhimillisen ymmärryksen ulottumattomissa. Valtavan opetusdatan takia ihmisen on käytännössä mahdotonta varmistaa myöskään datapohjan laatua. Näin ollen sekä toiminnan perusta että toimintalogiikka pakenevat inhimillisen hallinnan mahdollisuuksia.

378 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

379 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 (TiiS)*, *10*(4), 1–31.

380 Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. In *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm* (pp. 175–191). Cham: Springer International Publishing.

Toisinaan on argumentoitu, että indeterminanttien järjestelmien toiminnan laatua olisi mahdollista arvioida yksinkertaisempien algoritmien avulla. Tämä voi kuitenkin olla hyvin vaikeaa tai mahdotonta: jos ihmiskognitio ei kykene hahmottamaan testattavan järjestelmän toimintaa, on käytännössä mahdotonta varmistaa, toimiiko arvioiva algoritmi asianmukaisesti. Jos datan laatua ei voida kattavasti varmistaa eikä ennustemallin toimintaa ihmisjärjellä tai testaamisen avulla ymmärtää tai hallita, on jokseenkin odotettavaa, että järjestelmä tuottaa myös haitallista, harhaanjohtavaa tai valheellista materiaalia[381]. Esimerkiksi ChatGPT:n kohdalla on laajasti keskusteltu hallusinoinnista ja sen riskeistä[382]. Jos ja kun indeterministiset järjestelmät ovat suurelta osin äärimmäisen monimutkaisia, vaikeita tai mahdottomia tulkita, ja ne toimivat epälineaarisesti, järjestelmässä eri ominaisuuksista johtuvat riskien ennakoinnin ja hallinnan vaikeudet kertautuvat[383]. Hallitsemisen vaikeus ja siitä seuraava haittapotentiaali nousevat huomattavan korkeiksi ja riskit suuriksi[384].

Vaikka testaamalla ei voitaisikaan taata indeterminantin järjestelmän toiminnan täydellistä asianmukaisuutta, sen avulla pystytään havaitsemaan mahdollisesti huomattavakin osa riskeistä ja muokkaamaan järjestelmää niitä hillitseväksi. Niin testaamisen kuin käyttökokemusten perusteella järjestelmän ennustemallin päälle voidaan esimerkiksi lisätä erilaisia rajoitteita, joskin kaikkiin potentiaalisesti haitallisiin seikkoihin varautuminen on niidenkin avulla todennäköisesti mahdotonta. Epätäydellisyydestään huolimatta rajoitteet erilaisten mallien hallitsemiseksi ovat perusteltuja, sillä epätäydellinenkin haittapotentiaalin vähentäminen on luonnollisesti hyödyksi. Se kuitenkin pakottaa kysymään, kuinka suuria riskejä yhteiskunnassa ja globaalisti ollaan valmiita ottamaan uusien teknologisten työkalujen käytön mahdollistamiseksi.

Erityisesti generatiivisten, indeterminanttien järjestelmien tarjoamat mahdollisuudet esimerkiksi disinformaation tuottamiseen ja levittämiseen on tunnistettu jokseenkin laajasti[385]. Tällä hetkellä monet generatiiviset järjestelmät ovat avoimesti

381 Ks. esimerkiksi Gary, M. (2023). Inside the Heart of ChatGPT's Darkness (11.2.2023). *Marcus on AI.*

382 Ks. esimerkiksi Leiser, F., Eckhardt, S., Knaeble, M., Maedche, A., Schwabe, G., & Sunyaev, A. (2023). From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. *Proceedings of Mensch und Computer 2023* (81–90).

383 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

384 Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a Weapon of Mass Deception, Ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. *arXiv preprint arXiv:2304.11215*.

385 Ks. esimerkiksi De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120.

kenen tahansa käytettävissä. On siis melko varmaa, että tällaisilla järjestelmillä on merkitystä etenkin yksilön autonomiaan ja yhteiskunnan demokratiaan kohdistuvien seurannaishaittojen synnyssä. Mahdollisesta pahansuovasta käytöstä seuraavat haitat kiinnittyvät järjestelmien käytön tapoihin, jolloin niihin puuttuminen vaatisi käytön valvomista ja rajoittamista, mihin ei toistaiseksi ole järjestelmätasolla ryhdytty. Samankaltaista, haitallista toimintaa olisi toisaalta mahdollista harjoittaa myös ilman indeterminantin teknologian myötävaikutusta. Teknologia on tällöin työkalu, jonka ominaisuudet vaikuttavat haittojen syntymiseen lähinnä pahansuovan toimijan toimintamahdollisuuksia ja tehokkuutta kasvattamalla. Siinä toki järjestelmän ominaisuuksista johtuva monipuolisuus ja toimintamahdollisuuksien ennakoinnin ja hallinnan haasteet ovat merkityksellisiä tekijöitä.

Dynaamiset tekoälyjärjestelmät pystyvät oppimaan käytössä ollessaan eli niiden toiminnassa ympäristö vaikuttaa paitsi siihen, *millaisiin* lopputulemiin järjestelmä päätyy, myös siihen, *miten* se niihin päätyy. Käytännössä hyvin harva käytössä oleva järjestelmä oppii käytössä ollessaan. Koneoppivien järjestelmien oppiminen perinteisesti tapahtuu hallituissa olosuhteissa, joissa niiden kehittäjät pystyvät testaamaan niiden toimintaa ja varmistumaan sen asianmukaisuudesta ennen käyttöönottoa. Dynaamisten järjestelmien toimintalogiikka sen sijaan muuntuu käytön myötä. Koska toimintalogiikka muuttuu, toiminnan yksityiskohtaista logiikkaa ei pitkällä tähtäimellä pystytä kehitysvaiheessa varmistamaan. Käytännössä dynaamisten järjestelmien muuttuvien toimintatapojen asianmukaisuuden varmistaminen tarvitsee siis tuotantovaiheessa suunnitella siten, että järjestelmä ohjelmoidaan itsenäisesti testaamaan ja hyväksymään toimintalogiikkansa muutokset[386]. Ihmisen mahdollisuudet hallita järjestelmää siirtyvät siis niin ajallisesti kuin asiallisesti kauemmas ratkaisuista, joita järjestelmä käytössä ollessaan tekee.

Kun ihmisen tekemien suunnitteluaikaisten valintojen ja teknologian toiminnan välinen etäisyys kasvaa, myös ihmistoiminnan merkitys järjestelmän toiminnan määrittäjänä väistämättä heikkenee. Riskien tunnistaminen ja haittoihin varautuminen käyvät entistä vaikeammiksi. Riskit välittömille haitoille ovat ilmeisiä, sillä rajalliset hallinnan mahdollisuudet nostavat todennäköisyyttä odottamattomalle ja epätoivotulle toiminnalle.

Dynaamisten järjestelmien haittapotentiaalin kannalta ajan kuluminen on merkittävä tekijä. Ajan kuluessa järjestelmän toimintaympäristö muuttuu. On epätodennäköistä, että dynaamisten tekoälyjärjestelmien suunnittelijat pystyisivät suunnitteluvaiheessa varautumaan tuleviin, mahdollisesti nopeasti tapahtuviin muutoksiin niin algoritmisessa kuin fyysisessä ympäristössä. Tämä lisää haasteita riittävien suo-

386 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

jamekanismien suunnittelulle ja toteuttamiselle. Erilaisten teknologioiden vaikutuksia toisiinsa tai vaikutusten kumulaatiota voi olla mahdotonta ennakoida. Dynaamisten järjestelmien kohdalla yllättävillä tavoilla ja mahdollisesti nopeasti muuttuva ympäristö lisää riskiä hallitsemattomalle ketjureaktiolle ja niistä mahdollisesti seuraaville odottamattomille vaikutuksille, joita järjestelmään sisäänrakennettu muutosten varmistamisen menetelmä ei välttämättä tunnista. Tämä lisää riskiä paitsi välittömille haitoille myös mahdollisesti hallitsemattomaan algoritmiseen transformaatioon liittyville seurannaishaitoille.

Dynaamiset järjestelmät ovat haaste myös nykyiselle oikeusjärjestelmälle. Ihmiskeskeinen oikeusjärjestelmä lähtee siitä oletuksesta, että ainut oikeudellisesti merkityksellinen toimija on ihminen. Järjestelmät, joiden toiminnassa ihmisen tekemät valinnat muuttuvat jatkuvan oppimisen myötä koko ajan vähemmän merkittäviksi, poistavat ihmiselementin tavalla, johon oikeusjärjestelmää ei ole suunniteltu vastaamaan. Jos järjestelmä oppii toiminnassa ollessaan ja ilman ihmisen valvontaa, toiminnan tai sen seurausten palauttaminen ihmiseen on parhaimmillaankin keinotekoista. Samaan aikaan oikeusjärjestelmä ei – täysin perustellusti, kuten esimerkiksi Bryson ja muut[387] ansiokkaasti osoittavat – anna mahdollisuutta siirtää vastuuta toimijalle, joka ei ole ihminen[388].

Jos algoritmiset järjestelmät pystyvät toimimaan ja etenkin oppimaan ihmisen valinnoista ja hallinnasta aidosti irrallaan, päädymme tilanteeseen, jossa vastuu on mahdollista kohdentaa todelliseen toimijaan yhä harvemmin[389]. Tekoälyn tutkijat ovat puhuneet paljon vastuukuilusta (*responsibility gap*)[390]. Dynaamisten järjestelmien kohdalla kuilu syvenee, kun ihminen ei hallitse tai välttämättä lainkaan ymmärrä järjestelmän toiminnan muutoksia[391]. Kuilun syveneminen voi lisätä epäoikeudenmukaisuuden kokemuksia ja heikentää luottamusta yhteiskunnalliseen oikeu-

387 Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*, 273–291.

388 Tästä esimerkiksi Hallevy on eri mieltä. Ks. Hallevy, G. (2015). *Liability for crimes involving artificial intelligence systems*. New York, NY, USA: Springer International Publishing.

389 Jo 90–luvulla Solum katsoi, että tekoälytoimijoita ei tulisi määritellä oikeushenkilöksi muun muassa vastuukysymysten takia. Ks. Solum, L. B. (1991). Legal personhood for artificial intelligences. *NCL Rev.*, *70*, 1231.

390 Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

391 Vaikeudet vastuun kohdentamisessa nousevat esille myös ei-dynaamisten tekoälyjärjestelmien kohdalla. Esimerkiksi Coeckelbergh argumentoi: “[S]ociety deserves AI experts and operators who are in control, know what they are doing, and are *able and willing to communicate, explain and give reasons for what they are doing* to human and nonhuman moral patients.” Ks. Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4). 2066.

denmukaisuuteen. Mikäli näin tapahtuu, on todennäköistä, että myös yleinen luottamus[392] heikkenee. Yleinen luottamus taas vaikuttaa niin ihmisen edellytyksiin elää hyvää elämää kuin myös yhteiskunnan toimintamahdollisuuksiin. Järjestelmien dynaamisuus voi siis tulevaisuudessa olla yksi mahdollinen tekijä, joka lisää riskiä erilaisten laaja-alaisesti yhteiskuntaa muuttavien seurannaishaittojen ilmenemiselle.

Edellä läpikäytyjen, Viljasen hahmottelemien ominaisuuksien lisäksi myös monitai yleiskäyttöisyys on ominaisuus, joka nostaa huomattavasti järjestelmän haittapotentiaalia. EU:n tekoälysäädökseen lisättiin aivan viime metreillä yleiskäyttöiset tekoälymallit (*general purpose AI models*[393]), joita sääntelyn valmisteluvaiheessa kutsuttiin myös perustamalleiksi (*foundation model*[394]), jotta laajasta sovellettavuudesta johtuvat erityispiirteet voitaisiin sääntelyssä huomioida. Yleiskäyttöiset tekoälymallit ovat erilaisiin tekoälyjärjestelmiin liitettävissä olevia malleja, joita voidaan soveltaa erilaisissa konteksteissa eri tavoilla. Yleiskäyttöisiä tekoälymalleja ovat esimerkiksi kuvan- ja äänentunnistukseen luodut mallit sekä suuret kielimallit (*large language models*)[395].

Koska yleiskäyttöinen tekoälymalli voidaan ottaa käyttöön alkuperäisestä eroavissa konteksteissa, ensivaiheen tuottajan voi olla mahdotonta hahmottaa kaikkia relevantteja riskejä mallin suunnittelu- ja toteutusvaiheessa, saati pyrkiä varautumaan niihin. Tämä paitsi lisää mallin toiminnan ennakoimisen ja hallinnan haasteisiin liittyviä riskejä myös hämärtää vastuun muodostumista entisestään. Lisäksi riskit kertautuvat, kun mallia hyödynnetään erilaisten järjestelmien osana.

Monikäyttöiset, ihmisymmärryksen ja mahdollisesti luotettavan testaamismahdollisuuden ulottumattomissa[396] olevat tekoälymallit, joita pystytään hyödyntämään yhä uusissa konteksteissa, vaikuttavat väistämättä yhteiskuntaan. Osa muutoksista voi olla yhteiskunnallista todellisuutta syvästi mullistavia. On jopa esitetty, että

392 Rothstein, B., & Stolle, D. (2008). The state and social capital: An institutional theory of generalized trust. *Comparative politics*, *40*(4), 441–459.

393 Yleiskäyttöisiä tekoälymalleja säädellään EU:n tekoälysäädöksen luvussa V.

394 Ks. Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Art. 3(1c).

395 Ks. määritelmä: EU:n tekoälysäädös, artikla 3(63). Yleiskäyttöisiin tekoälymalleihin ja niiden sääntelyyn palataan luvussa 5.

396 Erilaisia mahdollisuuksia algoritmien hyödyntämiseen järjestelmien testaamisessa on pyritty kehittämään, ks. esimerkiksi Che, T., Liu, X., Li, S., Ge, Y., Zhang, R., Xiong, C., & Bengio, Y. (2021, May). Deep verifier networks: Verification of deep discriminative models with deep generative models. Konferenssijulkaisu, *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 35, No. 8, 7002–7010. Toisaalta on myös esitetty, että pyrkimyksistä huolimatta luotettava testaaminen ei ole mahdollista, ks. Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. Teoksessa *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm* (pp. 175–191). Cham: Springer International Publishing.

yleiskäyttöinen tekoäly on merkittävä askel kohti yleistekoälyä ja singulariteettia[397]. Tämä on kuitenkin toistaiseksi vasta hypoteettinen mahdollisuus. Vaikka singulariteetti jäisikin toteutumatta, on selvää, että mitä useampiin yhteyksiin yleiskäyttöisiä tekoälymalleja liitetään, sitä laajemmalle alueelle riski niiden mahdollisesti haitallisista vaikutuksista leviää.

Monikäyttöisistä järjestelmistä etenkin erilaista sisältöä tuottavat generatiiviset tekoälyjärjestelmät ovat viime aikoina herättäneet kiinnostusta. Ne ovat teknisiltä ominaisuuksiltaan lähtökohtaisesti itsenäisiä, monimutkaisia, tulkitsemattomia, epälineaarisia ja indeterminantteja – ja niitä voi käyttää monissa eri konteksteissa. Esimerkiksi tekstiä tuottavan ChatGPT:n on ehdotettu jo sopivan niin rikollisuuden torjuntaan, terapeuttiseen käyttöön kuin luovan työn avuksi huolimatta riskeistä, joita ominaisuuksiltaan erittäin riskialttiisiin teknologioihin väistämättä liittyy[398].

On selvää, että generatiiviset järjestelmät ovat monin tavoin mullistavia: monet niistä vaikuttavat ymmärtävän kieltä, sillä ne toimivat luonnollisella kielellä annettujen syötteiden avulla[399]. Niiden käyttäminen ei siis vaadi ohjelmointikoulutusta tai muuta tietotekniikan erikoisosaamista. Suurilla kielimalleilla ei kuitenkaan ole kykyä ymmärtää kielen suhdetta todellisen maailman ilmiöihin tai merkityksiin. Tietoteknisen erikoisosaamisen sijaan käyttäjiltä vaaditaan siis kykyä arvioida järjestelmien tuottamaa sisältöä kriittisesti. Tämä on paljon vaadittu, ja kuten aiemmin esiin nostettu Van Dijkin argumentaatio yhteiskunnan digitaalisesta jakautumisesta[400] osoittaa, kaikilla ei ole tällaiseen yhteneviä mahdollisuuksia. On todennäköistä, että kaikilla järjestelmiä käyttävillä ei ole riittävää ymmärrystä siitä, ettei järjestelmän generoima sisältö perustu moraaliseen arviointiin[401], tai ettei järjestelmä kykene erottamaan oikeaa väärästä tai totta valheesta. Jos käyttäjät eivät ymmärrä, minkälaisen teknologian kanssa he ovat tekemisissä, riskit voivat olla huomattavat[402] – etenkin,

397 Grossman, G. (2023). Generative AI may only be a foreshock to AI singularity. (11.2.2023). *VentureBeat.*

398 Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a Weapon of Mass Deception, Ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. *arXiv preprint arXiv:2304.11215*.

399 Taulli, T. (2023). Large Language Models: How Generative AI Understands Language. Teoksessa *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business* (93–125). Berkeley, CA: Apress.

400 Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

401 Tekoälyn mahdollisesta moraalisesta toimijuudesta ks. Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, *11*, 19–29.

402 Yhdysvalloissa käytiin toukokuussa oikeutta tapauksesta, jossa juristi oli hyödyntänyt ChatGPT:tä oikeustapausten hakemiseen. Oikeustapaukset, joita järjestelmä esitteli, eivät olleet todellisia. Ks. Heleskoski, J. (2023). Äärimmäisen nolo esimerkki tekoälyn vaaroista: emämunauksen tehnyt juristi perustelee tekoaan. (12.6.2023). *Mikrobitti*.

kun esimerkiksi ChatGPT:llä on taipumusta niin kutsuttuun hallusinoimiseen[403] eli todellisuuteen perustumattoman ”informaation” tuottamiseen. Ihmisten mahdollisuuksien ja kykyjen vaihdellessa generoivilla järjestelmillä on potentiaalia syventää digitaalista eriarvoistumista, jolloin mahdolliset haitat voivat kohdistua suhteettomasti yhteiskunnan haavoittuviin jäseniin.

Riskialttiiseen yhteyteen liitettynä haittapotentiaali luonnollisesti kasvaa: esimerkiksi suuret kielimallit voivat aiheuttaa vakavia haittoja, mikäli niitä käytetään esimerkiksi oikeudellisesti merkittävien päätösten luonnosteluun, puhumattakaan niiden hyödyntämisestä esimerkiksi hyvinvointipalveluiden tarjonnassa. Onkin selvää, että laajoista käyttömahdollisuuksista huolimatta kielimallien käyttöönotossa olisi syytä noudattaa ennemmin liikaa varovaisuutta kuin holtittomuutta. Toisaalta avoimesti tarjolla olevat sovellukset, kuten ChatGPT, vaikuttanevat jo laajasti toimintatapoihin niin yksilötasolla kuin monissa instituutioissakin. On epäselvää, miten tällaiset muutokset vaikuttavat esimerkiksi tiedonvälitykseen ja sitä kautta yksilöiden autonomiaan ja pidemmälle vietynä kulttuuriin, sivistykseen ja demokratiaan. Mahdollisten haittojen syntymiseen vaikuttanevat paitsi teknologian tekniset ominaisuudet ja niistä kumpuavat mahdollisuudet, myös esimerkiksi käyttäjän motiivit, teknologinen ymmärrys ja kyky arvioida teknologian toimintaa ja tuotoksia kriittisesti, sekä esimerkiksi eri instituutioissa mahdollisesti käyttöön otetut turvallisuusprotokollat.

Vaikka tekoälyteknologioiden ominaisuuksien suhde seurannaishaittoihin ei vaikuttaisi olevan suoraviivainen, voidaan arvioida, että riskiä lisäävät ominaisuudet lisäävät myös seurannaishaittojen muodostumisen mahdollisuuksia etenkin hallinnan mahdollisuuksien heikentymisen takia. Teknologioiden teknisiin ominaisuuksiin usein epäsuorasti kiinnittyvät seurannaishaitat tulisi ottaa vakavasti, sillä mitä laajemmin erilaisia monimutkaisia järjestelmiä liitetään yhteiskunnan eri tasoille ja rakenteisiin, sitä vahvemmin eri teknologioiden vaikutukset myös kumuloituvat ja kertautuvat, mikä voi aiheuttaa yllättäviä, vakavia seurauksia[404]. Haitat voivat ilmetä

403 Ks. esimerkiksi Petkauskas, V. (2023). ChatGPT’s answers could be nothing but a hallucination. (6.3.2023). *Cybernews*. Mahdottomuus erottaa tekoälyn tuottamia tekstejä, kuvia, videoita tai ääntä ihmisten tuottamista on omiaan lisäämään järjestelmien haittoja etenkin, kun järjestelmät pystyvät luomaan uskottavaa, mutta täysin valheellista materiaalia.

404 Myös Yeung: “[I]dentifying how multiple different algorithms might interact with other algorithmic agents in a complex and dynamic ecosystem generates risks of unpredictable, and potentially dangerous, outcomes. In other words, these interactions generate risks that we have barely begun to grasp.” Yeung, K. (2018). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *MSI-AUT (2018)*, *5.* 62.

kaukana teknologiasta, niin ajallisesti kuin maantieteellisestikin[405]. Algoritmisten teknologioiden ominaisuudet voivat paitsi vaikuttaa suoraan välittömien haittojen muodostumisen riskiin myös epäsuorasti lisätä mahdollisuuksia seurannaishaittojen ilmenemiselle, vaikeuttaa niiden havaitsemista[406], niiden syiden tunnistamista[407] ja niihin puuttumista.

Jotta algoritmisten teknologioiden haittapotentiaalia olisi mahdollista hallita, teknologioita koskevan lainsäädännön tulisi ensinnäkin tunnistaa potentiaaliset haitat ja toisekseen luoda tehokkaat prosessit niiden hallitsemiseksi. Seuraavassa luvussa käsitellään tekoälyyn liittyvää kansallista ja EU-tason lainsäädäntöä ja niiden kehitystä sekä arvioidaan, miten niissä tunnistetaan potentiaaliset haitat ja mahdollistetaan niiden rajoittaminen.

405 Teknologioiden leviämisestä sekä vaikutuksista ajassa ja paikassa, ks. myös McCarthy, D. R. (2013). Technology and 'the international' or: How I learned to stop worrying and love determinism. *Millennium*, *41*(3), 470–490.

406 Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, *89*, 1.

407 Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. 5.

5 Tekoälyn sääntely

Tekoälyn mukanaan tuomiin ongelmiin on havahduttu ympäri maailmaa, ja viime vuosina on eri konteksteissa nostettu esiin tarve tekoälyteknologioiden nimenomaiselle sääntelylle[408]. Tekoälyn sääntelyssä on kuitenkin törmätty erilaisiin haasteisiin. Uusia teknologioita kehitetään niin nopeasti, että säädännäisen oikeuden keinoin on vaikea reagoida riittävällä vauhdilla[409]. Sääntelypyrkimyksiä haastavat myös tekoälyteknologioiden ylikansallisuus ja pelko siitä, että liian rajoittava sääntely heikentää valtioiden globaaleja vaikutusmahdollisuuksia sekä ylikansallista markkina- ja valta-asemaa[410].

Tekoälysääntelyn haasteet eivät luonnollisesti ole ainutlaatuisia: uusien teknologioiden sääntelystä on käyty teknologisen kehityksen vanavedessä laajasti keskustelua[411]. Monet uudet innovaatiot historian saatossa ovat tuoneet mukanaan riskejä ja haittoja, jotka on kuitenkin ymmärretty vasta ajan kuluessa. Esimerkkeinä mainittakoon erilaiset ajoneuvot, monet lääkkeet ja geenimanipulaatio[412]. Kun modernia yhteiskuntaa sääntelevät kiihtyvyyden lainalaisuudet, potentiaalisten riskien ja haittojen tunnistamiselle jää yhä vähemmän aikaa, jolloin myös sääntely käy yhä haastavammaksi. Esimerkiksi Jasanoff on perustellusti kysynyt, miten demokratia voidaan ylläpitää teknologioiden kyllästämässä maailmassa, jossa tieto rakentuu erikoisosaamisen varaan, arvot ovat ristiriitaisia, tuottajat toimivat kaukana kuluttajista ja loppukäyttäjistä, ja yhteisen kielen ja normien puute estää yhtenäiskulttuurin olemassa-

408 Ks. esimerkiksi Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Springer.

409 Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. Teoksessa *Temporal Boundaries of Law and Politics* (72–87). Routledge.

410 Smuha, N. A. (2021). From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, *13*(1), 57–84.

411 Ks. esimerkiksi Koulu, A. R. (2018). Digitalisaatio ja algoritmit – oikeustiede hukassa? *Lakimies*, *116*(7–8), 840–867.

412 Jasanoff, S. (2020). Constitutional moments in governing science and technology. Teoksessa *The Ethics of Nanotechnology, Geoengineering, and Clean Energy* (pp. 477–494). Routledge. 622.

olon[413]. Kuten luvussa 1.2 esitin, Rosa kuvaa, kuinka yhteiskunnallisen kiihtyvyyden taustalla vaikuttavat kolme moottoria, joiden toiminta johtaa kiihtyvään muutokseen: taloudellinen, rakenteellinen ja kulttuurinen moottori. Näistä rakenteellinen moottori edesauttaa toimintojen – ja sitä kautta myös tiedon, arvojen ja ideologioiden – eriytymistä[414]. Argumentaatio linkittyy Jasanoffin kysymykseen ja nostaa esiin modernin maailman ajallisten lainalaisuuksien aikaansaamat, yhä laajenevat haasteet demokraattisen päätöksenteon kontekstissa.

Teknologioiden sääntelyssä näkyy sama problematiikka. Lainsäädännön ja sitä toteuttavan viranomaiskoneiston yksi kenties olennaisimmista tehtävistä on turvata yhteiskunnan vakaus ja ennakoitavuus[415]. Globaalissa maailmassa, jossa on väistämättä huomioitava paitsi yhä monimutkaistuva lähiympäristö myös monipolviset ja ylikansalliset vaikutussuhteet, merkityksellisten seikkojen määrä kasvaa valtavasti. Samalla yhä kapea-alaisemmat erikoistumisalueet rajoittavat niin päättäjien kuin kansalaistenkin mahdollisuuksia hahmottaa laajoja kokonaisuuksia ja erilaisten vaikutusten verkostoja. Tämä tekee lainsäädäntötyöstä jatkuvasti vaikeampaa: kun ympäristö muuttuu ja monimutkaistuu kiihtyvällä vauhdilla, tarvitaan yhä enemmän ja yhä nopeammin tehtyjä poliittisia ja lainsäädännöllisiä päätöksiä. Näin ollen tarpeellisten päätösten määrä kasvaa samalla, kun aikaikkuna niiden tekemiselle kaventuu ja päätösten tekemiseksi tarpeellisen tiedon määrä lisääntyy[416]. Toisaalta myös päätösten vaikutukset kasvavat: jos päätetään sääntelyn suunnasta, sen vaikutukset kertautuvat ajassa ja tilassa yhä nopeammin. Käytännössä tästä seuraa, että kokonaiskuvan hahmottamisen vaikeutuminen yhdessä päätösten vaikutusten kasvamiseen vaatisivat poliittiselle päätöksenteolle ja sääntelyn toteuttamiselle *enemmän* aikaa, kun taas kiihtyvän yhteiskunnan aikapaine *vähentää* aikaa, joka päätöksiin on mahdollista käyttää. Tarve uusille, tehokkaammille apuvälineille ja päätöksentekotyökaluille kasvaa – ja tähän tarpeeseen pyritään vastaamaan algoritmisten teknologioiden avulla.

Nopeasti muuttuvien ilmiöiden, kuten juuri erilaisten teknologisten innovaatioiden, mukanaan tuomat mahdollisuudet yhdistettynä sääntelemisen haasteisiin ovatkin tuoneet poliittiseen ja myös akateemiseen keskusteluun monia uusia näkökulmia lainsäädännön, teknologioiden ja yhteiskunnan vuorovaikutussuhteista[417]. Lainsää-

413 Ibid.

414 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

415 Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. Teoksessa *Temporal Boundaries of Law and Politics* (72–87). Routledge. 77.

416 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 262.

417 Ks. esimerkiksi Bijker, W. E., & Law, J. (Eds.). (1994). *Shaping technology/building society: Studies in sociotechnical change*. MIT press; ja

Jones, M. L. (2018). Does technology drive law? The dilemma of technological exceptionalism in cyberlaw. *U. Ill. JL Tech. & Pol'y*, 249.

dännön voi nähdä toisaalta osana yhteiskunnallista järjestelmää ja siten – rakenteistumisteorian termistöllä puhuttaessa – osana rakennetta, joka määrittelee yhteiskunnallista todellisuutta. Toisaalta se on yhteiskunnan toimijoiden luomaa, ja laajasti ajatellen sitä voidaan pitää teknologiana, jonka ihmiset ovat rakentaneet yhteiskunnallisen järjestyksen ylläpitämiseksi. Lainsäädännöstä voikin puhua yhteiskunnan vakauttajana (*stabilizer*) tai jopa *hidastajana (decelerator),* joka parhaimmillaan hillitsee liian nopeasti kiihtyvää muutosta[418]. Lainsäädännön kaltaisesti myös tekoälyteknologiat määrittelevät modernin maailman toiminnalle reunaehtoja ja muovaavat yhteiskunnallisia järjestelmiä[419], jolloin myös niillä on vaikutusta siihen, millaisiksi yhteiskunnalliset käytänteet muodostuvat. Algoritmisten teknologioiden kehitystä tai käyttöä ei kuitenkaan ohjaile tavoite vakaudesta ja ennakoitavuudesta vaan ennemmin pyrkimys uuden luomiseen, tehostamiseen ja muutokseen. Samalla ne toimivat yhteiskunnallisen todellisuuden kiihdyttäjinä (*accelerator*)[420].

Algoritmisten teknologioiden ja lainsäädännön rajapintaan liittyviä teemoja voi lähestyä monista eri suunnista. Aihepiirin puitteissa voidaan arvioida esimerkiksi, miten algoritmisten teknologioiden avulla voidaan tehostaa hallintoa (*governance by algorithms*[421], *algorithmic regulation*) tai luoda lainkaltaista sääntelyä (*technological management*[422]), tai miten algoritmisia teknologioita voidaan ja/tai halutaan säännellä lain keinoin tai muuten (*governance/regulation of algorithms*).

Kuten toisessa osatutkimuksessa tuotiin esiin, vaikuttaa siltä, että inhimillisen toimintanopeuden ja -tehokkuuden jäädessä riittämättömäksi modernin maailman kiihtyessä, algoritmiset järjestelmät vaikuttavat paitsi houkuttelevilta, myös välttämättömiltä, jotta toimintojen tehostaminen voidaan taata[423]: *algoritmisen sääntelyn* avulla voisi olla mahdollista kiihdyttää toimintoja, jotka manuaalisesti toteutettuna vaikuttavat väistämättä liian hitailta nykytilanteessa. Samaan aikaan *algoritmisten teknologioiden ja tekoälyn sääntely* kuitenkin on vasta työn alla, ja kiihdyttäminen tapahtuu vailla lainsäädännön ohjailua, mikä altistaa huomattaville algoritmiselle haitoille. Joissakin yhteyksissä, kuten Suomessa nähtiin ennen hallintolain uudistusta, lainsäädännön puuttuminen rajoittaa algoritmisten teknologioiden hyödyntä-

418 Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. Teoksessa *Temporal Boundaries of Law and Politics* (72–87). Routledge. 72.

419 Brownsword, R. (2015). In the year 2061: from law to technological management. *Law, Innovation and Technology*, *7*(1), 1–51.

420 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

421 Ks. esimerkiksi Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1–18.

422 Ks. esimerkiksi Brownsword, R. (2016). Technological management and the rule of law. *Law, Innovation and Technology*, *8*(1), 100–140.

423 Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707. 12.

mistä ja sitä kautta niistä seuraavia mahdollisia riskejä ja haittoja. Mikäli näin käy, samalla eritahdistumiseen eli muun yhteiskunnan vauhdista jäämiseen liittyvät, tehottomuudesta kumpuavat analogiset haitat korostuvat[424].

Tässä luvussa käsittelen tekoälyn ja algoritmisten teknologioiden sääntelyä kansallisessa lainsäädännössä ja EU:n tasolla. Keskityn tekoälyteknologioiden nimenomaiseen sääntelyyn, joten käsittelyn ulkopuolelle jäävät esimerkiksi tietoturva- ja datasääntely, vaikka ne totta kai ulottavat vaikutuksensa myös algoritmisten teknologioiden alueelle.

Ensimmäisessä alaluvussa käsitellään kansallista sääntelyä. Suomessa algoritmisten teknologioiden käyttöä säännellään toistaiseksi vain julkishallinnossa, minkä takia arvioin hallinnollisen päätöksenteon automaation sääntelyä ja siihen liittyviä lainvalmistelumateriaaleja. Toisessa alaluvussa siirrytään käsittelemään tekoälyn sääntelypyrkimyksiä EU:ssa. Pääpaino on EU:n tuoreessa tekoälyasetuksessa, jota taustoitan avaamalla kehityskulkuja, jotka EU:ssa edelsivät tekoälyasetusehdotuksen julkaisemista.

5.1 Kansallinen sääntely

Suomen lainsäädännössä tekoälyteknologioita ei säännellä nimenomaisesti. Tekoälyn hyödyntämistä hallinnollisessa päätöksenteossa on kuitenkin rajoitettu merkittävästi, sillä kun käytetään tekoälyteknologioita, harkintavallan käyttöä on vaikea kontrolloida: on vaikea varmistua siitä, että päätöksenteossa noudatetaan hyvän hallinnon periaatteita tai, että oikeusturva ja virkavastuu toteutuisivat. Vaikka sääntely ei kohdistu nimenomaisesti tekoälyteknologioihin vaan niiden käyttöön rajatussa kontekstissa, katson tarpeelliseksi kuvata sitä tilannetta, jossa kansallisesti olemme nyt. Painotus on työn teeman mukaisesti niissä seikoissa, joiden erityisesti voi nähdä luovan riskejä yhteiskunnallisten, algoritmisten haittojen syntymiselle.

Suomessa tekoäly nousi kansallisessa politiikassa vahvasti esille viimeistään, kun Sipilän hallituksen hallitusohjelma 2015 määritti oikeistolaisessa hengessä tavoitteeksi, että digitalisaation avulla otettaisiin tuottavuusloikka. Hallitusohjelmassa tavoiteltiin Suomen ketterää uudistamista muun muassa sääntelyä purkamalla, hallinnollista taakkaa keventämällä ja kannustamalla kokeilukulttuurin käyttöönottoon[425].

Hallitusohjelma johti Tekoälyaika-hankkeen käynnistymiseen. Hankkeen ensimmäinen raportti julkaistiin 2017, ja siinä asetettiin tavoitteeksi, että Suomi nos-

[424] Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 270–291.

[425] Hallitusohjelma 2015. 26.

tettaisiin tekoälyn soveltamisen kärkimaaksi[426]. Viimeinen, kokoava raportti, jota yleisesti kutsutaan Suomen tekoälystrategiaksi[427], julkaistiin vuonna 2019. Tekoälystrategiassa korostettiin yhtäältä ihmislähtöisen tekoälyn merkitystä ja kansalaisten osallistamista, toisaalta yritysten vastuuta eettisten ohjeiden luomisessa ja itsesääntelyn toteuttamisessa. Laintasoisen sääntelyn kehittämistarvetta strategiassa käsiteltiin heikosti.

Kuten osatutkimuksessa 3[428] argumentoidaan, valtion politiikassa on vahvasti rohkaistu hyödyntämään tekoälyn mahdollisuuksia. Politiikkaratkaisuilla on pyritty kannustamaan siihen, että tekoälyä otettaisiin käyttöön mahdollisimman laajasti. Niin kutsuttu kokeilukulttuuri nähtiin jo vuoden 2015 hallitusohjelmassa keinona tekoälyratkaisujen monipuolisten mahdollisuuksien käyttöönottamiseksi[429] – sen opportunistisesta ja jokseenkin oikeusvaltioperiaatteelle vieraasta painotuksesta huolimatta. Tämä innosti tekoälyn käyttöönottoon niin julkisella kuin yksityiselläkin sektorilla. Julkishallinnon automaatiopyrkimykset törmäsivät kuitenkin nopeasti perustuslaillisiin esteisiin.

Perustuslakivaliokunta (PeV) otti laajasti kantaa automaattisen päätöksenteon mahdollisuuksiin julkishallinnossa lausunnossaan 62/2018 ja myöhemmin lausunnossaan 7/2019. Lausunnot olivat vastauksia kahdelle hallituksen esitykselle (HE 224/2018 ja HE 18/2019), joilla pyrittiin mahdollistamaan Maahanmuuttoviraston päätöksenteon osittainen automatisointi. PeV katsoi, että automaattiseen päätöksentekoon liittyy ”sellaisia perusoikeuksiin ja julkisen vallan käyttöön liittyviä erittäin laajakantoisia oikeudellisia erityiskysymyksiä, joita pitäisi arvioida yleislainsäädännön kehittämistarpeiden kautta”[430]. Samalla PeV katsoi, että ”on välttämätöntä pidättäytyä uusista hallinnonalakohtaisista automatisoitua päätöksentekoa koskevista sääntelyehdotuksista”[431] ennen kuin tarve yleislaintasoiselle sääntelylle on selvitetty.

426 Työ- ja elinkeinoministeriö. (2017). Suomen tekoälyaika: Suomi tekoälyn soveltamisen kärkimaaksi: Tavoite ja toimenpidesuositukset. Työ- ja elinkeinoministeriön julkaisuja 41/2017. 16.

427 Työ- ja elinkeinoministeriö. (2019). *Edelläkävijänä tekoälyaikaan: Tekoälyohjelman loppuraportti.* Työ- ja elinkeinoministeriön julkaisuja 2019:23.

428 Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 275; Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

429 Ks. Hallitusohjelma 2015. 26; Hallitusohjelma 2019. 107;
Työ- ja elinkeinoministeriö. (2017). Suomen tekoälyaika: Suomi tekoälyn soveltamisen kärkimaaksi: Tavoite ja toimenpidesuositukset. Työ- ja elinkeinoministeriön julkaisuja 41/2017. 28; ja
Työ- ja elinkeinoministeriö. (2019). *Edelläkävijänä tekoälyaikaan: Tekoälyohjelman loppuraportti.* Työ- ja elinkeinoministeriön julkaisuja 2019:23. 87–88.

430 PeVL 7/2019. 12.

431 PeVL 7/2019. 11.

Perustuslakivaliokunnan lausunto 7/2019 merkitsi käytännössä automaattisen päätöksenteon kieltoa julkishallinnossa, ja siitä lähti liikkeelle Arviomuistion (Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista)[432] valmistelutyö. Arviomuistio valmistui heinäkuussa 2020, ja sen lopputulemana esitettiin, että ”sallittu automaatio tulisi rajata tilanteisiin, joissa ratkaisu on johdettavissa koneellisesti lainsäädännöstä ja tiedossa olevista yksiselitteisistä faktoista tilanteessa, jossa päätöksentekoon ei liity harkintavaltaa”[433]. Päätöksenteko voisi siis perustua ”vain viranomaisen ennalta määrittelemiin, lainsäädännön mukaisiin päättelysääntöihin”[434]. Arviomuistiossa ehdotettiin, että virkavastuu tulisi kohdentaa automaattisen päätöksentekojärjestelmän käyttöä koskevien virkatehtävien perusteella[435]. Jokseenkin samanlaista ratkaisua ehdotettiin HE 224/2018:ssa[436], mutta PeV:n mukaan se ei riittänyt täyttämään perustuslaillisia vaatimuksia[437]. Erona voidaan kuitenkin nähdä arviomuistiossa esitetty kolmen kohdan lista virkatoimista, joista vastaaviin henkilöihin virkavastuu tulisi kohdentaa: 1) järjestelmässä käytettävien päätöksentekosääntöjen hyväksyminen, 2) järjestelmän tai siihen tehtyjen muutosten hyväksyminen ja vaatimustenmukaisuuden arviointi sekä 3) järjestelmän valvonta[438].

Arviomuistiosta järjestettiin lausuntokierros heti sen julkaisemisen jälkeen, ja lausunnonantajien joukosta valittiin työryhmä valmistelemaan hallituksen esitystä. Hallituksen esitys 145/2022 julkaistiin 19.9.2022, ja eduskuntakäsittelyn jälkeen se vahvistettiin muutettuna 23.3.2023. Lakimuutokset astuivat voimaan 1.5.2023. Hallintolakiin lisättiin uusi luku, 8b, jossa määritellään automaattisen ratkaisun edellytykset (53e §), oikeussuojaedellytykset (53f §), sekä vaatimukset automaattisesta ratkaisusta ilmoittamiselle (53g §). Virkavastuun kohdentumista automaattisen päätöksenteon kontekstissa säädellään laissa julkisen hallinnon tiedonhallinnasta (906/2019), johon lisättiin säädökset automaattisen päätöksentekojärjestelmän toimintaprosessista. Prosessi pitää sisällään vaiheet järjestelmän kehittämisestä, käyttöönotosta ja toiminnan seuraamisesta. Kullekin vaiheelle on nimettävä henkilö, joka vastaa siitä, että menettely on asianmukaista[439]. Algoritmiseen päätöksentekoon kyt-

432 Vainio, N., Tarkka, V., & Jaatinen, T. (2020). *Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista*. Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2020:14.

433 Ibid. 63.

434 Ibid. 63.

435 Ibid. 64.

436 HE 224/2018. 41.

437 PeVL 62/2018. 8.

438 Vainio, N., Tarkka, V., & Jaatinen, T. (2020). *Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista.* Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2020:14. 64–65.

439 HE 145/2022. 174–178.

keytyvään virkavastuuseen liittyy vaikeita kysymyksiä, joiden tutkiminen on tärkeää. Aihepiirin laajuuden vuoksi sen käsittely on kuitenkin rajattu tämän työn ulkopuolelle[440].

Yksinkertaiset, sääntöpohjaiset algoritmiset teknologiat ovat hallintolain uuden 8b luvun valossa tietyin ehdoin sallittuja hallinnollisessa päätöksenteossa. Päätösautomaatio ja päätöksenteon tukena hyödynnettävät avustavat teknologiat muuttavat väistämättä hallintolainkäytön reunaehtoja. Hyvän hallinnon periaatteissa toimivalta kuuluu viranomaiselle laissa säädetyin edellytyksin. Jos viranomaisen harkinta- ja toimivaltaa ohjailevat lainsäädännön edellytysten lisäksi data ja siitä jalostettu informaatio sekä koodi, lähtökohta väistämättä muuttuu[441]: virkamiehen harkintavalta rajautuu järjestelmän toimintamuodon ja merkitykselliseksi arvioidun datan perusteella. Mahdollisesti tästä seuraavat ongelmat lienee ainakin jossain määrin tunnistettu lainsäädännössä, sillä päätösautomaatiossa on sallittu vain rajatusti algoritmisten teknologioiden käyttö. Sallitun automaation muodot noudattelevat pitkälti Arviomuistion linjaa. Automaattisten päätöksentekojärjestelmien käyttöala on rajattu päätöksiin, jotka eivät vaadi tapauskohtaista harkintaa tai harkinta on toteutettu ennakollisesti:

> Viranomainen voi ratkaista automaattisesti asian, johon ei sisälly seikkoja, jotka edellyttävät tapauskohtaista harkintaa, tai johon sisältyvät tapauskohtaista harkintaa edellyttävät seikat virkamies tai muu asian käsittelijä on arvioinut. Ratkaisemisen on perustuttava sovellettavan lain ja etukäteisen harkinnan perusteella laadittuihin julkisen hallinnon tiedonhallinnasta annetun lain (906/2019) 2 §:n 16 kohdassa tarkoitettuihin käsittelysääntöihin.

Lainsäädännössä sallitaan ainoastaan sääntöpohjaiset järjestelmät. Oppivat tekoälyteknologiat on suljettu pois lopullisen hallinnollisen päätöksen teosta. Sääntöpohjaisten järjestelmien käyttöala on kuitenkin haluttu venyttää laajaksi: kuten lainauksesta näkyy, lainsäädännössä mahdollistetaan harkinnanvaraisten seikkojen *ennakollinen arvioiminen*. Se luo perustan myös niin sanottujen tyyppitapausten automaattiselle käsittelylle. Hallituksen esityksessä viranomaiselle annetaan jokseenkin laajat mahdollisuudet tunnistaa tyyppitapauksia:

440 Ks. aiheeseen liittyen esimerkiksi Hirvonen, H. (2022). Virkavastuu ja päätösautomaatio–vastuun henkilökohtaisuus kriisissä? *Lakimies*, *2022*(3–4), 386–418.

441 Koivisto, I., & Koulu, R. (2020). Miten hyvä hallinto digitalisoidaan? Haaste oikeustieteelliselle tutkimukselle. *Lakimies*, *118*(6), 798–821;
Pöysti, T. H. (2018). Kohti digitaalisen ajan hallinto-oikeutta. *Lakimies*, *2018*(7–8), 868–903.

> … viranomaisen on tehtävä etukäteinen harkinta siitä, mitkä asiat ovat luonteeltaan sellaisia, että ne voidaan ratkaista automaattisesti, ja laadittava käsittelysäännöt, joiden perusteella asiat valitaan ja ratkaistaan. Käytännössä viranomaisen on tunnistettava käsittelemiensä asioiden joukosta olennaisilta piirteiltään samanlaisina toistuvat asiat (*tyyppitapaukset*), jotka viranomaisen arvion mukaan tulee ratkaista aina saman säännön mukaisesti.
>
> Yksinkertaisimmillaan tyyppitapaus voidaan tunnistaa suoraan sovellettavan aineellisen lainsäädännön perusteella, kun lainsäädäntö on riittävän yksiselitteistä. Tilanteissa, joissa aineellinen lainsäädäntö soveltuu siinä käytettyjen käsitteiden perusteella laajasti erilaisiin tapauksiin, viranomaisen on tyyppitapauksia tunnistaessaan otettava ennakolta kantaa myös sovellettavien säännösten tulkintaan. Myös asioissa, joissa päätösten perustana oleva lainsäädäntö sisältää runsaasti tulkinnanvaraisuutta (esimerkiksi liikennevakuutuslaki ja sen taustalla oleva vahingon- korvauslaki), voi olla mahdollista tunnistaa tyyppitapauksia, joissa viranomaisella (tai lakisääteisten vakuutusten osalta julkista hallintotehtävää hoitavalla yksityisellä) on jossakin asiaryhmässä vakiintunut ratkaisukäytäntö, jonka mukaan tiettyjen piirteiden mukaiset asiat ratkaistaan aina samalla tavalla. Tyyppitapaus voi myös perustua raja-arvojen tarkasteluun esimerkiksi siten, että tietyn euromääräisen rajan ylittävien tulojen katsotaan tietynlaisissa tilanteissa aina täyttävän laissa vaaditun edellytyksen[442].

Tyyppitapauksiksi voidaan siis määritellä tapausryhmät, joissa käytäntö on riittävän vakiintunut. Tällöin katsotaan toteutunutta käytäntöä. Kuten Koivisto ja Koulu argumentoivat, hallinnollinen päätöksenteko on, tai ainakin sen usein tulisi olla, päämääräorientoitunutta[443] ja siinä pyritään valitsemaan toimintatapa, joka oikeassa suhteessa tavoiteltuun päämäärään nähden, kuten hallintolain 6 § vaatii. Käytännössä hallinnollisessa päätöksenteossa näkyy – tai tulisi näkyä – lainsäädännössä ja sitä ohjaavissa periaatteissa rakentuva, historiallisesti luotu sosiaalipoliittinen tavoitteellisuus[444]. Mikäli ratkaisu tehdään ennakollisen harkinnan perusteella, se on luonteeltaan täysin erilaista: päämäärän sijaan arvioidaan vastaavuutta aiempiin tapauksiin ja pyritään toistamaan aiempia käytänteitä. Vaikka tyyppitapausten kohdalla tulkintakäytännön tulee olla vakiintunut ja harkintavallan soveltamisen yhdenmukaista, jotta tällaisia tapausryhmiä voidaan laillisesti automatisoida, on riskinä, että automatisointi jähmettää käytännön. Tällöin viranomaisharkinnan merkitys heikkenee ja

442 HE 145/2022. 98.
443 Koivisto, I., & Koulu, R. (2020). Miten hyvä hallinto digitalisoidaan? Haaste oikeustieteelliselle tutkimukselle. *Lakimies*, *118*(6), 798–821. 801.
444 Ibid. 812.

järjestelmän, datan ja koodin merkitys kasvaa. Jos päätöskohtainen harkinta muuttuu mekaaniseksi sääntöjen soveltamiseksi eikä käytäntöä arvioida säännöllisesti, mahdolliset muutostarpeet huomattaneen vasta huomattavalla viiveellä. Automaattinen prosessi myös väistämättä vie mahdollisuuden huomioida tyyppitapauksista yllättävällä tai odottamattomalla tavalla poikkeavat tilanteet, jotka ihmiskäsittelijä todennäköisesti pystyisi ottamaan huomioon. Tämä tuo hallinnolliseen päätöksentekoon massaharkinnan elementtejä[445]. Ennakkoon toteutettu päätösharkinta vaikuttanee voimakkaimmin niin kutsuttuihin rajatapauksiin, joiden kohdalla riski oikeudenloukkauksille voi kasvaa[446].

Potentiaalisista ongelmista huolimatta lainsäädännössä sallitun päätösautomaation lisäämät riskit ja haitat jäänevät melko rajatuiksi. Lainsäädäntö ei ainoastaan lisää sallitun automaation määrää vaan vaatii myös jo käytössä pitkään olleen automaation saattamista lain sallimiin rajoihin, mitä voi pitää erinomaisena. Toki oletettavissa on, että automaatio tulee lisääntymään nyt, kun sen reunaehdoista on oikeudellinen varmuus[447]. Automaattisen päätöksenteon lisääntymisen seuraukset selviävät vasta käytänteiden vakiintuessa.

Toisin kuin automaattisten päätösten tekemistä, hallintolaissa ei rajoiteta avustavassa roolissa olevien teknologioiden käyttöä. Niin kauan kuin ihminen tekee lopullisen päätöksen, mitä vain tekoälyjärjestelmiä voidaan käyttää esimerkiksi päätöksen luonnosteluun[448]. Mahdollisuus käyttää myös esimerkiksi monimutkaisiin ennustemalleihin perustuvia tai tulkitsemattomia tekoälyjärjestelmiä avustavassa roolissa voi tehostaa päätöksentekoprosesseja, mutta samaan aikaan vaikuttaisi ilmeiseltä, että huomattava määrä valtaa voi siirtyä teknologisille järjestelmille. Vaikutukset voivat olla merkittäviä. Ihmisten alttiutta luottaa liikaa automaattisiin järjestelmiin, niin kutsuttu automaatiovinoumaa (*automation bias*), on tutkittu viime vuosina laa-

445 Myös Koivisto ja Koulu huomauttavat, että massaluonteisten hallintopäätösten vaatimukset ovat muun muassa hyvän hallinnon periaatteiden osalta täysin samat kuin kaikkien muidenkin hallintopäätösten. Näin ollen päätösvolyymistä riippumatta yksilöllinen harkinta tulee taata myös automaattisessa päätöksenteossa. Ks. ibid. 816.

446 Kun huomioidaan myös esimerkiksi kritiikki, jota Maahanmuuttoviraston massaharkinta (collectivized discretion) vuoden 2015 niin kutsutun pakolaiskriisin jälkeen kohtasi, voidaan kysyä, johtaako tyyppitapausten määrittely massaharkinnan laintasoiseen mahdollistamiseen, ja mikäli näin, minkälaisissa tilanteissa. Hallinnollisen päätöksenteon tulisi kuitenkin olla aina tapauskohtaista, ja nähtäväksi jää, miten tyyppitapausten ottaminen päätöksenteon pohjaksi vaikuttaa tämän toteutumiseen käytännössä. Ks. Vanto, J., Saarikkomäki, E., Alvesalo-Kuusi, A., Lepinkäinen, N., Pirjatanniemi, E., & Lavapuro, J. (2022). Collectivized Discretion: Seeking Explanations for Decreased Asylum Recognition Rates in Finland After Europe's 2015 "Refugee Crisis". *International Migration Review*, *56*(3), 754–779.

447 Koivisto, I. (2023). Automaattinen päätöksenteko tulee – oletko valmis? (2.3.2023). *Perustuslakiblogi: Suomen valtiosääntöoikeudellisen seuran ajankohtaispalsta.*

448 HE 145/2022. 96.

jasti. Ei ole täysin selvää, missä tilanteissa ja kuinka vahvasti automaatiovinouma vaikuttaa, mutta on esitetty perusteltu huoli siitä, että teknologioiden käyttö päätöksenteon apuna heikentää päätöksien taustalla olevaa harkintaa ja sitä kautta lopullisten päätösten laatua[449]. Liiallinen luottaminen teknologioihin altistaa siten virheellisille ratkaisuille sekä lisää riskiä syrjivälle tai epäoikeudenmukaiselle päätöksenteolle. Järjestelmien käytöstä mahdollisesti juontuvat muutokset eivät myöskään välttämättä ole järjestelmien käyttäjien eivätkä päätöksen kohteen havaittavissa tai ymmärrettävissä.

Vaikka lainsäädännössä vastuu kohdennetaan päätöksen tehneeseen viranhaltijaan, on selvää, että inhimillinen päätöksenteko muuttuu, jos tekoälyjärjestelmiä käytetään viranomaistyön apuna[450]. Mikäli vaikeasti ymmärrettäviä, monimutkaisiin ennustemalleihin perustuvia tekoälyjärjestelmiä otetaan laajasti käyttöön päätösten tekemisen tueksi, on mahdollista, että tekoälyjärjestelmät ohjaavat päätöksentekijöitä painottamaan menneitä käytänteitä päämäärän kustannuksella. Monissa kansainvälisissä tutkimuksissa on tunnistettu tekoälyjärjestelmien vaikutukset päätöksentekoon[451]. Mitä suuremmassa roolissa viranomaistyötä avustava tekoälyjärjestelmä on, sitä voimakkaampi vaikutus sillä todennäköisesti on lopputuloksen muodostumisessa.

Kun laki säätelee ainoastaan lopullisten hallintopäätösten automatisointia, jää viranomaisten oman harkinnan varaan, milloin ja minkälaisia avustavia järjestelmiä otetaan käyttöön ja kuinka laajasti niitä hyödynnetään. Samaan aikaan jatkuvasti lisääntyvä paine hallinnon tehostamiselle[452] kannustaa tehokkaampien järjestelmien käyttöönottoon, minkä seurauksena yhä suurempi määrä hallinnollista työtä ja myös päätösharkintaa siirtynee lähitulevaisuudessa algoritmisten järjestelmien tekemäksi. Tällä on väistämättä laaja-alaisia seurauksia yhteiskunnassa. Hallinnollinen valta ja siihen liittyvät kysymyksenasettelut muodostuvat yhä suuremmassa määrin osaksi algoritmisten järjestelmien suunnittelukysymyksiä[453], ja toisaalta algoritmisten jär-

449 Ks. esimerkiksi Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge.

450 Ks. esimerkiksi Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance*, *2*(4), 301–313.

451 Ks. esimerkiksi Bahl, U., Topaz, C. M., Obermüller, L., Goldstein, S., & Sneirson, M. (2023). Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. *UCLA L. Rev. Discourse*, *71*, 246.

452 Ks. Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 270–291.

453 Pöysti, T. H. (2018). Kohti digitaalisen ajan hallinto-oikeutta. *Lakimies*, *2018*(7–8), 868–903.

jestelmien suunnitteluratkaisut määrittävät yhä suuremmassa määrin hallinnollisen päätöksenteon ja harkintavallan reunaehtoja.

Avustavan automaation sääntelemättömyys voi lisätä haittojen riskiä merkittävästi. Kun viranomaisilla on mahdollisuus hyödyntää tekoälyteknologioita avustavassa roolissa ilman lainsäädännössä asetettuja rajoitteita, harkintavalta tosiasiallisesti muuttunee aiempaa enemmän dataan perustuvaksi ja algoritmisten järjestelmien määrittämien reunaehtojen mukaiseksi. Näin ollen riskit tekoälyjärjestelmien opetus- ja muuhun dataan palautuville haitoille kasvanevat, samoin kuin haitat, joita harkintavallan tosiasiallinen kaventuminen voi aiheuttaa. Käytännössä tämä tarkoittaa virheellisten ja vinoutuneiden, tosiasiallisesti algoritmisten järjestelmien toimintaan palautuvien päätösten sekä mahdollisesti niiden ilmentämän epäoikeudenmukaisuuden seurauksena potentiaalisesti syntyviä välittömiä ja seurannaishaittoja, jotka voivat paitsi kohdistua yksilöön myös kokonaisiin ryhmiin ja yhteiskuntaan aiemmin tässä johdanto-osiossa kuvatuin tavoin.

Suomessa on odotettu EU:n tekoälyasetusta määrittelemään tarkemmin, minkälaisia algoritmisia teknologioita saadaan käyttää ja missä yhteyksissä. Tämä lienee syy sille, että kansallinen lainsäädäntö on vielä toistaiseksi kapea-alaista, eikä nimenomaista tekoälysääntelyä juurikaan ole. Seuraavaksi käsittelen EU:n tekoälysääntelyä. Tuore tekoälysäädös on jokseenkin jännitteisessä suhteessa kansallisen sääntelyn kanssa: kansallinen sääntely hallinnollisen päätöksenteon automaatiosta vaikuttaa olevan huomattavasti EU:n lainsäädäntöä rajoittavampaa. Hallintolaissa sallittua automaatiota ei EU:n tekoälysäädöksessä ylipäätään lueta tekoälyksi, mikä tarkoittaa, että tekoälyasetus ei vaikuta kohdistuvan siihen. Toisaalta on epäselvää, onko kansallisesti mahdollista rajoittaa tekoälyn hyödyntämistä asetusta tiukemmin. Mikäli ei, mahdollisuudet tekoälyn hyödyntämiseen myös hallinnollisessa päätöksenteossa laajentunevat nykyisestä.

5.2 EU-sääntely

EU:ssa tekoälyn sääntelyä on valmisteltu pitkään, ja tekoälyasetusehdotusta edelsi laaja työ tekoälykysymysten parissa. EU:n tekoälystrategia[454] julkaistiin vuonna 2018. Siinä määriteltiin keskeisiksi tavoitteiksi tekoälyn käytön lisääminen ja mahdollistaminen, valmistautuminen sosioekonomisiin muutoksiin sekä asianmukaisen eettisen ja oikeudellisen kehyksen käyttöönoton varmistaminen[455].

Tekoälyä käsittelevä korkean tason asiantuntijaryhmä (High-Level Expert Group on AI) perustettiin tekoälystrategian tavoitteiden mukaisesti, ja sen työsken-

454 Euroopan komissio. (2018.) Tekoäly Euroopassa.
455 Ibid. 3.

telyn tuloksena julkaistiin *Luotettavaa tekoälyä koskevat eettiset ohjeet*[456] huhtikuussa 2019. Ohjeissa tekoälyn tärkeimmiksi ominaisuuksiksi nostettiin lainmukaisuus, eettisyys ja luotettavuus. Ohjeissa tekoälyn potentiaaliset riskit ymmärrettiinkin laajasti, ja tekoälyn tuottajia kehotettiin huomioimaan myös yllättävät vaikutukset esimerkiksi ”demokratiaan, oikeusvaltioperiaatteeseen ja oikeudenmukaiseen jakautumiseen tai itse ihmismieleen”[457]. Ohjeissa keskityttiin kuitenkin nimenomaisesti eettisyyteen ja luotettavuuteen, ja lainsäädäntöön liittyvät kysymykset jätettiin niitä huomattavasti pienempään rooliin[458].

Vuonna 2020 EU:n *Valkoinen kirja tekoälystä – Eurooppalainen lähestymistapa huippuosaamiseen ja luottamukseen*[459] julkaistiin. Valkoinen kirja esitteli riskiperusteisen sääntelymallin, mikä todennäköisesti on toiminut tekoälysäädösehdotuksen lähtökohtana. Valkoisessa kirjassa riskien katsottiin kiinnittyvän toisaalta tekoälyjärjestelmän ominaisuuksiin ja toisaalta sen käyttöympäristöön. Sen sijaan eettisissä ohjeissa esiin nostettuja vaikutuksia demokratiaan, oikeusvaltioperiaatteeseen, oikeudenmukaisuuteen tai ihmismieleen ei valkoisessa kirjassa enää käsitelty, vaan tekoäly kuvattiin työkaluna, jolla voi olla ”merkittävä rooli kestävän kehityksen tavoitteiden saavuttamisessa ja demokraattisen prosessin ja sosiaalisten oikeuksien tukemisessa”[460]. Ennakoimattomista riskeistä[461] tai yllättävistä vaikutuksista kirjassa ei ole mainintoja. Vaikuttaakin siltä, että valkoisen kirjan myötä näkökulma siirtyi painottamaan helposti tunnistettavia riskejä, joita tekoäly voi aiheuttaa ihmisoikeuksille ja turvallisuudelle.

Vuoden 2021 huhtikuussa komissio julkaisi ehdotuksen EU:n tekoälyasetukseksi[462]. Säädöksen lopullinen sisältö varmistui maaliskuussa 2024, kun parlamentti äänesti viimeisenä trilogi-kompromissin hyväksymisestä. Säädös julkaistiin EU:n virallisessa lehdessä 12.7.2024 ja astui voimaan elokuussa 2024, ja siirtymäaikojen jälkeen säännökset tulevat sovellettaviksi portaittain: säännökset kielletyistä tekoälyjärjestelmistä kuuden kuukauden, yleiskäyttöisiä tekoälymalleja koskevat

456 Tekoälyä käsittelevä korkean tason asiantuntijaryhmä. (2019). Luotettavaa tekoälyä koskevat eettiset ohjeet.

457 Tekoälyä käsittelevä korkean tason asiantuntijaryhmä. (2019). Luotettavaa tekoälyä koskevat eettiset ohjeet. 16.

458 ”Ohjeissa ei käsitellä nimenomaisesti luotettavan tekoälyn ensimmäistä osa-aluetta (lainmukainen tekoäly), vaan niillä pyritään pikemminkin antamaan opastusta toisen ja kolmannen osa-alueen (eettinen ja luotettava tekoäly) edistämiseksi ja varmistamiseksi.” ibid. 8.

459 Euroopan komissio. (2020). Valkoinen kirja tekoälystä.

460 Ibid: 2.

461 Valkoinen kirja kyllä mainitsee tekoälyjärjestelmien mahdollisuuden ennakoimattomaan toimintaan, mutta yhdistää sen luomat riskit ainoastaan vaikeuteen varmistaa perusoikeuksien toteutumisen valvonta. Ks. ibid: 13.

462 Euroopan komissio. 2021/0106 (COD). Proposal for Artificial Intelligence Act.

velvoitteet 12 kuukauden, ja suuririskisiä järjestelmiä koskevat velvoitteet joko 24 kuukauden tai 36 kuukauden kuluttua voimaan astumisesta.

Säädös on täysharmonisoiva ja jättää jäsenvaltioille hyvin vähän liikkumavaraa tekoälyyn liittyvissä kysymyksissä. Tarkoituksena on varmistaa, että EU:n alueella tuotettavat, markkinoitavat ja/tai käytettävät tekoälytuotteet ovat turvallisia ja kunnioittavat ihmisoikeuslainsäädäntöä ja EU:n arvoja (art. 1). Sääntelyn piiristä kuitenkin on rajattu pois yksinomaan jäsenvaltioiden kansalliseen turvallisuuteen ja maanpuolustukseen sekä ainoastaan tutkimustarkoituksiin kehitetyt ja käytetyt tekoälysovellukset (art. 2).

Sääntelyratkaisussa päädyttiin määrittämään tekoälyn tuotannon, markkinoille saattamisen ja käyttöön ottamisen vaatimukset niihin liitetyn riskitason mukaan. Riski voi liittyä tekoälyjärjestelmän ominaisuuksiin tai sen käyttökontekstiin, ja riskitaso vaikuttaa siihen, minkälaisia velvoitteita tekoälyjärjestelmien tuottajien/tarjoajien ja käyttöönottajien tulee noudattaa. Pieni määrä tekoälysovelluksista päätettiin kieltää kokonaan, minkä lisäksi riskikategorioita on kolme: suuri, rajattu ja pieni.

Säädös koskee ensisijaisesti tekoälyjärjestelmiä, jotka määritellään artiklassa 3(1):

> ‘AI system’ means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;

Tekoälyjärjestelmien lisäksi säädöksessä säännellään yleiskäyttöisiä tekoäly*malleja*, joista on asetusehdotuksen kehityskaaren aikana käytetty myös nimitystä perustamallit. Yleiskäyttöiset tekoälymallit määritellään artiklassa 3(63):

> “‘general purpose AI model’ means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable to competently perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications. This does not cover AI models that are used before release on the market for research, development and prototyping activities.”

Tekoälyjärjestelmien riskitasoon perustuvan sääntelyn lisäksi säädöksessä määrätään erillisiä velvoitteita yleiskäyttöisten tekoälymallien tarjoajille. Mallin tarjoajan tulee täyttää luvussa V säädetyt dokumentointi-, testaus-, ja tiedonantovelvoitteet, joiden tarkempi sisältö määräytyy mallin teknisten ominaisuuksien perusteella.

Lisäksi luvussa IV säädetään myös yleiskäyttöisiä tekoälymalleja koskevista läpinäkyvyysvaatimuksista. Kun yleiskäyttöinen tekoälymalli liitetään osaksi tekoälyjärjestelmää tai siihen liitetään esimerkiksi käyttöliittymä, jolloin se täyttää (yleiskäyttöisen) tekoälyjärjestelmän määritelmän, järjestelmiä koskevat riskiperusteiset velvoitteet kohdistuvat myös siihen. Riskitasoon sidottujen velvoitteiden täyttämisestä vastaa tekoälyjärjestelmän tarjoaja, joka voi olla joko mallin tuottaja tai joku muu taho.

EU:n tekoälysäädöksellä pyritään puuttumaan riskeihin, joita tekoälysovellukset voivat aiheuttaa ihmisten terveydelle, turvallisuudelle ja perus- ja ihmisoikeuksien toteutumiselle. Tämän nimissä osa tekoälysovelluksista on asetuksen voimaantulon jälkeen EU:ssa kiellettyjä. Näistä säädetään artiklassa 5.

Säädöksen 6 artiklassa määritellään luokittelusäännöt suuririskisille tekoälysovelluksille. Artiklan ensimmäisen kohdan mukaan suuren riskin sovelluksia ovat ensinnäkin tekoälytuotteet ja tällaisten tuotteiden turvallisuuskomponentit silloin, kun ne kuuluvat jonkin liitteessä I listatun EU:n harmonisoivan sääntelyn alaan[463], ja joiden kohdalla tulee olemassa olevan sääntelyn perusteella suorittaa ulkopuolinen arviointiprosessi tuotteen vaatimustenmukaisuudesta. Harmonisoitua sääntelyä on EU:ssa huomattava määrä, ja sen perusteella suuririskisiä ovat esimerkiksi osa leluihin, koneisiin ja lääkinnällisiin laitteisiin liitettävistä tekoälyjärjestelmistä. Näiden lisäksi, 6 artiklan toisen kohdan mukaisesti, suuririskisiksi lukeutuvat tekoälyjärjestelmät, joita käytetään liitteessä III listattuihin tarkoituksiin, joihin palaan myöhemmin tässä luvussa.

Suuririskisiin sovelluksiin liittyvät velvoitteet on määritelty laajoiksi. Suuren riskin tekoälysovelluksiin kohdistuvat vaatimukset kuvataan III luvun 2 osassa, ja sovelluksen tarjoajaan ja käyttöönottajaan (joka viittaa esimerkiksi viranomaiseen, joka hyödyntää sovellusta toiminnassaan, eikä siis loppukäyttäjään) kohdistuvista velvoitteista säädetään III luvun 3 osassa. Lisäksi säädöksessä määrätään esimerkiksi jäsenvaltioiden viranomaisten velvoitteista (luku III osa 4), suuririskisten tekoälyjärjestelmien standardoinnista, sertifioinnista ja rekisteröinnistä (luku III osa 5) sekä markkinoille laskettujen tekoälyjärjestelmien valvonnasta (luku IX).

Tekoälyjärjestelmien kolmas riskikategoria pitää sisällään rajatun riskin sovellukset, joita koskee ainoastaan rajallinen määrä läpinäkyvyysvaatimuksia (artikla 50), jotka asettavat velvoitteen muun muassa merkitä selkeästi, mikäli esimerkiksi kuva- tai videosisältö on tuotettu tekoälyn avulla. Suurin osa tekoälysovelluksista kuuluu neljänteen, pienen riskin kategoriaan, eikä niihin kohdisteta erityisiä vaatimuksia.

463 Liitteessä on eroteltu osa A, joka sisältää sääntelyn, joka on harmonisoitu sillä perusteella, että se kuuluu New Legislative Frameworkin alaan, sekä osa B, jossa listataan muulla perusteella harmonisoitu sääntely.

EU:n tekoälyasetus vaikuttaa ensisilmäyksellä kunnianhimoiselta. Eri riskikategoriat vaikuttavat huomioivan tekoälysovellusten erilaiset piirteet kattavasti ja asettavan niiden tuottamiselle, markkinoille saattamiselle ja käytölle perusteltuja rajoituksia. Tarkemmin katsottaessa näyttää kuitenkin siltä, että rajoitteet jäävät todellisuudessa melko keveiksi, ja täysharmonisoivana säädös vaikuttaa itseasiassa luovan jäsenvaltioille velvollisuuden sallia laaja-alainen tekoälyn tuottaminen, markkinointi ja hyödyntäminen[464].

5.2.1 Kielletyn riskin sovellukset

Säädöksessä esitetty kielletyn riskin sovellusten lista on lyhyt, ja siinä kielletään manipuloivat ja ihmisten haavoittuvaisuuksia hyväksikäyttävät järjestelmät, tietyt sosiaalisen pisteytyksen järjestelmät, profilointiin perustuvien tekoälyjärjestelmien käyttäminen rikosten riskin arviointiin ja ennakolliseen valvontaan sekä tällaisten järjestelmien markkinoille saattaminen, kasvojentunnistustietokantojen luomiseen tai laajentamiseen tarkoitettujen tekoälysovellusten saattaminen markkinoille ja käyttäminen, tunteiden tunnistamiseen työpaikoilla ja oppilaitoksissa tarkoitettujen tekoälysovellusten saattaminen markkinoille ja käyttäminen, biometrinen luokittelu tiettyjen ominaisuuksien mukaan, sekä julkisilla paikoilla tapahtuva reaaliaikaisten biometristen tunnistamisjärjestelmien käyttäminen viranomaistyössä. Tiukat vaatimukset ja lukuisat poikkeukset hapertavat sääntelyn vakuuttavuutta.

Komission alkuperäisessä ehdotuksessa kiellettyjä käyttötapoja oli neljä: manipuloivat ja haavoittuvuuksia hyödyntävät järjestelmät, sosiaalisen pisteytyksen järjestelmät ja reaaliaikainen biometrinen tunnistaminen viranomaistyössä lukuun ottamatta sallittuja poikkeustilanteita. Ehdotus keräsi paitsi kiitosta myös kritiikkiä muun muassa suppeudestaan[465] ja kiellettyihin tai suuren riskin järjestelmiin liittyvien säädösvaatimusten mielikuvituksellisuutta lähentyvistä rajoituksista[466]. Lopullisessa versiossa kiellettyjen sovellusten listaus on kattavampi, mutta toisaalta säädökseen lisätty rajoitus, joka sulkee säädöksen ulkopuolelle kaikki valtion kansalli-

464 Näin katsovat myös: Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112.

465 Ks. esimerkiksi Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an artificial intelligence act. *SSRN 3899991*.

466 Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112.

seen turvallisuuteen ja maanpuolustukseen liittyvät käyttötarkoitukset (art. 2(3)), vesittää jossain määrin kieltojen merkitsevyyttä.

Manipuloivien järjestelmien kiellosta säädetään artiklan 5 kohdissa 1 (a) ja (b), joista ensimmäinen koskee järjestelmiä, jotka hyödyntävät ”subliminaalisia” eli alitajuntaan vaikuttavia tekniikoita, ja jälkimmäinen järjestelmiä, jotka hyväksikäyttävät yksilön haavoittuvaista asemaa, joka liittyy määriteltyihin ominaisuuksiin.

Kiellot on kiinnitetty vaatimukseen siitä, että järjestelmä aiheuttaa tai voi aiheuttaa (”causes or is reasonably likely to cause”) merkittävää haittaa joko manipuloinnin kohteelle tai toiselle henkilölle. Parlamentin esittämän mukaan[467] lopullisessa säädöstekstissä haitta on määritelty resitaalissa 5 kattamaan sekä aineellisen että aineettoman haitan, jotka pitävät sisällään fyysisen, psyykkisen, yhteiskunnallisen ja taloudellisen haitan. Määritelmän ulkopuolelle jäävät esimerkiksi ihmisen autonomiaan kohdistuvat haitat[468], jotka tulevat arvioiduiksi todennäköisesti vain, mikäli niistä seuraa henkilöön kohdistuvia fyysisiä, psyykkisiä tai taloudellisia haittoja.

Säädöksen mukaan haitan tulee aiheutua joko siitä, että järjestelmä hyödyntää alitajuntaan vaikuttavia tekniikoita (”subliminal techniques”) tai muita tekniikoita, jotka heikentävät ihmisen mahdollisuuksia tehdä tietoisia päätöksiä (”informed decision”), tai siitä, että se käyttää hyväksi ihmisen tai ihmisryhmän haavoittuvaa asemaa. Resitaalin 29 mukaan haitta voi syntyä, jos järjestelmä joko hyödyntää erilaisia komponentteja tavoilla, jotka ylittävät ihmisen havainnointikyvyn, tai harhauttaa muutoin ihmisen toimimaan oman tahtonsa vastaisesti (”such AI systems deploy subliminal components such as audio, image, video stimuli that persons cannot perceive, as those stimuli are beyond human perception, or other manipulative or deceptive techniques that subvert or impair person’s autonomy, decision-making or free choice in ways that people are not consciously aware of those techniques or, where they are aware of them, can still be deceived or are not able to control or resist them.”). Tämän voi resitaalin mukaan saavuttaa esimerkiksi virtuaalitodellisuuden tai kone-aivo-rajapinnan keinoin. Vaatimus muodostuu kohtuullisen korkeaksi, osin jopa mielikuvitukselliseksi[469].

467 Parlamentti lienee huomannut haitan määrittelyn puutteellisuuden, sillä vasta parlamentin ehdotuksessa haitta määriteltiin ensimmäistä kertaa. Lopullinen muotoilu vastaa parlamentin ehdottamaa. Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Muutos 14.

468 Ks. myös Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: a response to the European Commission’s proposal for an artificial intelligence act. *SSRN 3899991*. 21.

469 Ks. myös Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 98–99.

Vaikka resitaalin 5 mukaisesti haitta voi ulottua yhteiskuntaan, manipuloivien tekoälyjärjestelmien kiellossa sitä ei nosteta esiin. Vaikuttaisikin siltä, että mikäli manipuloivan tekoälyjärjestelmän aiheuttama haitta kohdistuu esimerkiksi oikeusjärjestelmään tai demokratiaan, kielto ei välttämättä ulottuisi siihen. Toisaalta resitaalissa 29 huomioidaan myös ajan kuluessa kumuloituvat haitat, joihin suuri osa yhteiskuntaan kohdistuvista haitoista lukeutunee. Resitaalissa kuitenkin huomautetaan, ettei tekoälyjärjestelmän tarjoajan tai käyttöönottajan voida olettaa toimivan tarkoituksellisesti, mikäli käytöksen vääristyminen johtuu (”results from”) tekoälyjärjestelmään liittymättömistä seikoista, jotka ovat tarjoajan tai käyttöönottajan ennakointi- ja toimintamahdollisuuksien ulottumattomissa. Jokseenkin ristiriitaisesti samassa resitaalissa tosin jatketaan, että tarkoitusta tuottaa huomattavaa haittaa ei tekoälyjärjestelmän tarjoajalta tai käyttöönottajalta vaadita, mikäli haitta on seurausta tekoälyjärjestelmän mahdollistamista manipuloivista käytänteistä. Ilmeisesti siis tarjoajalta tai käyttöönottajalta vaaditaan tarkoitusta muokata ihmisen käyttäytymistä, mutta ei välttämättä tarkoitusta tuottaa sillä haittaa, vaan riittää, että manipulointi tosiasiallisesti johtaa tai voi johtaa merkittävän haitan syntymiseen. Manipulointia, joka määritelmällisesti tarkoittaa ihmisen ohjailua usein vasten tämän tahtoa manipuloijan haluamaan suuntaan, ei siis itsessään nähdä vielä haitallisena, vaan sen todelliset tai mahdolliset seuraukset ratkaisevat.

Säädöksen seuraava kielto (art. 5(1) alakohta (c)) kohdistuu sosiaalisiin pisteytysjärjestelmiin, ja sen nojalla kiellettyjä ovat tekoälyjärjestelmät, joita käytetään luonnollisten henkilöiden tai ryhmien arviointiin tai luokitteluun käyttäytymisen tai tiedettyjen, tulkittujen tai ennustettujen henkilökohtaisten ominaisuuksien tai persoonallisuuden piirteiden perusteella. Alkuperäisessä komission ehdotuksessa kielto kohdennettiin ainoastaan julkisvallan käyttäjiin tai heidän puolestaan toimiviin, mutta lopullisesta sanamuodosta rajaus poistettiin, ja kielto koskee siis myös yksityisiä toimijoita. Laajennus on tervetullut.

Säännöksen sanamuoto on kuitenkin varsin mielenkiintoinen: sosiaalisen pisteytyksen järjestelmät eivät sen mukaan vaikuttaisi olevan kiellettyjä, elleivät ne johda (”leading to”) määriteltyihin lopputuloksiin. Kieltoa ei ole sidottu itse järjestelmään tai edes sen käyttöön vaan järjestelmän avulla luodun sosiaalisen pisteytyksen seurauksiin. Esimerkiksi manipuloivien järjestelmien kohdalla asetuksessa säädetään järjestelmä kielletyksi, mikäli se aiheuttaa tai *voi aiheuttaa* merkittävää haittaa. Sosiaalisen pisteytyksen järjestelmien kohdalla tällainen epätoivotun seurauksen *mahdollisuuteen* kiinnittyvä kieltävä elementti puuttuu.

Kielto on myös rajattu kohdistumaan käyttäytymisen sekä tiedettyjen, tulkittujen tai ennustettujen henkilökohtaisten ominaisuuksien ja persoonallisuudenpiirteiden pisteyttämiseen. Mikä kaikki näiden alueelle katsotaan kuuluvaksi, jää epäselväksi. Esimerkiksi sijaintitiedot voivat kertoa epäsuorasti henkilön ominaisuuk-

sista, mutta niiden perusteella pisteyttäminen ei vaikuttaisi olevan säädöksen mukaan kiellettyä [470].

Riskiarviointijärjestelmien kieltoa (art. 5 (1) alakohta (d)) ei komission alkuperäisessä ehdotuksessa ollut, mutta sen päätyminen lopulliseen säädöstekstiin on erinomainen asia. Kiellon perusteella ihmisten algoritminen profilointi ei saa olla ainut tapa arvioida yksilöiden riskiä syyllistyä rikoksiin. Toisaalta profilointia ja riskiarviointia saa säädöksen mukaan käyttää inhimillisen arvioinnin tukena, mikä voi jättää kiellon merkityksen lopulta varsin vähäiseksi.

Myös kasvojentunnistustietokantojen luomisen ja laajentamisen kielto (alakohta (e)) kohdentamattomalla haravoinnilla internetistä tai valvontakamerakuvista lisättiin säädökseen viimemetreillä, kuten myös tunteidentunnistusjärjestelmien kielto työ- ja opiskeluympäristöissä (alakohta (f)). Lisäykset vahvistavat säädöstä merkittävästi. Erityisesti kasvojentunnistustietokantojen luomisen ja laajentamisen kieltoa kuitenkin heikentää se, että kansallisen turvallisuuden alaan kuuluvat asiat jäävät säädöksen ulkopuolelle. Toisaalta EU:n oikeuskäytännön mukaan kansallinenkaan turvallisuus ei voi perustella mitä vain EU-oikeuden vastaisia toimia[471]. Lienee kuitenkin todennäköistä, että kansallisen turvallisuuden nimissä tuotettavien ja käytettävien, tekoälysäädöksen vastaisten tekoälyjärjestelmien sallittavuutta tullaan selvittämään EU-oikeudessa tulevina vuosina.

Artiklakohdan 5(1) alakohdassa (g) kielletään biometrisen datan perusteella luokittelu, kun tarkoituksena on päätellä esimerkiksi yksilön poliittinen kanta, uskonto tai seksuaalinen suuntautuminen, ja alakohdassa (h) reaaliaikainen, etäältä tapahtuva biometrinen tunnistaminen lainkäytössä. Lainkäytön (”for the purpose of law enforcement”) määritelmä vastaa esitutkintaviranomaisten toimia (artikla 3(46)). Reaaliaikaisen biometrisen tunnistamisen kieltoon on liitetty mittava lista poikkeustilanteita, jolloin säännöstä voidaan poiketa. Poikkeukset koskevat tiettyjen vakavien rikosten uhrien ja kadonneiden henkilöiden etsimistä, vakavien henkeen tai terveyteen kohdistuvien uhkien ja terroristisen rikosten estämistä ja rikollisten tunnistamista ja paikantamista silloin, kun etsityn epäillyn tekemästä rikoksesta langetettava maksimirangaistus on vähintään neljä vuotta vankeutta ja se löytyy asetuksen liitteen II listalta. Näitä tarkoituksia varten muutoin kiellettyjä biometrisen tunnistamisen järjestelmiä voidaan käyttää, jos se on välttämätöntä (”strictly necessary”). Tämän rajaarvon määrittely vaikuttaisi jäävän esitutkintaviranomaisten tai näiden toimia valvovan viranomaisen arvioinnin varaan. Jotkut kansalaisjärjestöt ovat esittäneet huolen,

[470] Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: a response to the European Commission’s proposal for an artificial intelligence act. *SSRN 3899991*. 23.

[471] Euroopan ihmisoikeustuomioistuimen tutkimusosasto. (2013). *National security and European case-law*. Euroopan neuvosto/EHCR.

että poikkeukset vesittävät kiellon merkityksen ja luovat riskin järjestelmien väärinkäytölle ja siitä seuraaville oikeuksien loukkauksille[472]. Riski voi olla olemassa erityisesti, jos järjestelmän käytöstä päätetään rikosepäilyn perusteella, vailla kattavia tietoja rikoksesta ja sen tekijästä.

5.2.2 Suuren riskin sovellusten sääntelykehikko

Tekoälyjärjestelmistä vain pieni osa määrittyy suuren riskin sovellukseksi siksi, että se sisältyy liitteen I harmonisoidun sääntelyn alueelle tai liitteen III kahdeksan kohdan listaan. Resitaalissa 52 avataan, milloin tekoälyjärjestelmä, joka ei ole artiklan 6(1) mukainen NLF-tuote tai sen turvallisuuskomponentti, on suuririskinen. Tällaisiksi on määritelty järjestelmät, joihin liittyy suuri riski haitalle, joka kohdistuu yksilöiden terveyteen, turvallisuuteen tai perusoikeuksiin. Haitan suuruutta arvioitaessa on huomioitavat sekä haitan vakavuus että todennäköisyys. Lisäksi olennaista on, että järjestelmää käytetään artiklassa 6(2) viitatun liitteen III listan mukaiseen käyttötarkoitukseen. Kahdeksan kohdan listalle on otettu muun muassa julkisten palveluiden ja avustusten kohdistamiseen liittyviä käyttötarkoituksia, kuten myös kriittiseen infrastruktuuriin, rekrytointiin, koulutukseen, oikeuslaitokseen, oikeuden toimeenpanemiseen, maahanmuuttoon ja biometriseen tunnistamiseen liittyviä käyttötarkoituksia.

Suuren riskin järjestelmien tuottajille määrätään varsin laajat toimintavelvoitteet järjestelmien laadun varmistamiseksi (art. 17) ja riskien hallitsemiseksi (art. 9). Säädöksen II luvun toisessa ja kolmannessa osassa määrätään muun muassa dokumentoinnista ja lokitietojen keräämisestä, opetusdatan laadusta, läpinäkyvyydestä, ihmisen suorittamasta valvonnasta, järjestelmien vakaudesta, virheettömyydestä ja kyberturvallisuudesta. Toisaalta resitaalissa 66 painotetaan myös tarvetta välttää perusteettomia rajoituksia markkinoille. Näennäisesti laajoja velvoitteita heikennetään myös lukuisilla rajoituksilla. Esimerkiksi datan laatukriteereistä todetaan artiklassa 10(3) seuraavasti: “Training, validation and testing data sets shall be relevant, representative, and to the best extent possible, free of errors and complete in view of the intended purpose.” Neuvoston kannassa vaatimustaso määriteltiin hyvin matalalle, kun resitaalissa 44[473] (lopullisessa säädöstekstissä datan laatua käsitellään resitaa-

472 Ks. esimerkiksi Day, J., Iwańska, K., Simon, E. & Willamo, K. (2024). *Packed with loopholes: Why the AI act fails to protect civic space and the rule of law.* Civil Liberties Union for Europe e.V.

473 Eurooppa-neuvosto. General approach on EU AI Act. (14954/22). Resitaali 44: “These datasets should also be as free of errors and complete as possible in view of the intended purpose of the AI system, taking into account, in a proportionate manner, technical feasibility and state of the art, the availability of data and the implementation of appro-

lissa 67) katsottiin, että vaatimukset tulisi suhteuttaa muun muassa datan saatavuuteen, ja tarvittaessa riskienhallinnan toimilla voitaisiin kompensoida opetusdatan laadun puutteita. Näin matalaa tasoa ei kuitenkaan sementoitu säädöstekstiin. Vaatimustaso jäänee silti varsin vaatimattomaksi, sillä datan on oltava ainoastaan riittävän kuvaavaa (”sufficiently representative”), ja tekoälyjärjestelmän tarjoajan on otettava käyttöön sopivat (”appropriate”) datanhallintamenetelmät (resitaali 67). Riittävyyden ja sopivuuden arviointi vaikuttaisi jäävän tarjoajan vastuulle.

Suuririskisten tekoälyjärjestelmien osalta vasta parlamentti esitti, että sääntelyssä velvoitettaisiin tekoälyjärjestelmien käyttöönottajat ihmisoikeusvaikutusten arvioimiseen. Velvoite lisättiinkin lopulliseen säädöstekstiin, joskin kevennettynä parlamentin ehdottamasta. Artiklassa 27 asetetaan suuren riskin tekoälyjärjestelmän käyttöönottajalle velvoite arvioida ihmisoikeusvaikutuksia siinä kontekstissa, jossa järjestelmää olisi tarkoitus käyttää, mikäli käyttöönottaja on julkisoikeudellinen taho tai tarjoaa julkisia tai niihin rinnastettavia palveluita. Julkispalveluihin rinnastuvia palveluita voivat olla esimerkiksi pankkitoiminta ja henki- ja terveysvakuutusten tarjoaminen. Näissä käytettävät, tekoälyä hyödyntävät riskiarviointi- ja hinnoittelujärjestelmät tulee siis arvioida ihmisoikeusvaikutusten varalta. Käytännössä velvoite on rajattu koskemaan erittäin pientä osaa suuren riskin tekoälyjärjestelmien käyttöönottajista.

Ihmisoikeusvaikutusten arvioinnissa on oltava tiedot siitä, keihin tai mihin ryhmiin järjestelmän käyttöönoton arvioidaan vaikuttavan ja minkälaisia haittoja näihin voi kohdistua. Lisäksi on kuvattava, miten ihminen valvoo järjestelmän toimintaa (human oversight), ja esitettävä toimintasuunnitelma riskien toteutumisen varalle. Verrattuna parlamentin esittämään vaikutusarviointiin lopullinen velvoite jää vaatimattomaksi ja rakentuu pitkälti inhimillisen valvonnan mahdollistamisen ja toteuttamisen – johon liittyy lähes yli-inhimillisiä elementtejä[474] – (art. 14) varaan. Alkuperäinen parlamentin esittämä velvoite koski ensinnäkin kaikkia suuririskisten tekoälyjärjestelmien käyttöönottajia, ei pelkästään julkishallinnollisia tai niihin vertaantuvia toimijoita. Lisäksi siinä olisi vaadittu, että arvioinnissa on mukana, ”to the best extent possible”, myös niiden ryhmien tai tahojen edustajia, joihin tunnistetut riskit kohdistuvat. Tällaisista edustajista mainittiin esimerkiksi tasa-arvoelimet ja kuluttajansuojaviranomaiset. Lopullisessa säädöstekstissä, vaikka ihmisoikeusvaikutusten arvioinnista siinä säädettiinkin, ei vaadita lainkaan ulkopuolisten tahojen mukanaoloa.

priate risk management measures so that possible shortcomings of the datasets are duly addressed.”

474 Ks. inhimillisen valvonnan yli-inhimillisistä vaatimuksista, I., Koulu, R., & Larsson, S. (2024). User accounts: How technological concepts permeate public law through the EU’s AI Act. *Maastricht Journal of European and Comparative Law,* 1023263X241248469.

Käytännössä siis vaikutusten arviointi ja riskien hallinta jätetäänkin pääosin suuren riskin tekoälyjärjestelmien tarjoajien vastuulle, kun taas käyttöönottajien vaikutustenarviointivelvoitteet jäävät asetuksessa melko kapea-alaisiksi. Valinta kohdentaa riskien hallinnan ja minimoinnin velvoitteet niille tahoille, joilla on paras ymmärrys tekoälyjärjestelmän ominaisuuksista, mutta toisaalta väistämättä käyttöönottajaa heikompi ymmärrys ympäristöstä, jossa järjestelmä tulee käytettäväksi. Suuren riskin tekoälyjärjestelmien tarjoajille myönnetään varsin laaja harkintavalta asianmukaisten ja kohdennettujen (”appropriate and targeted”) riskinhallintatoimien päättämiseksi, jotta riskien vähentämisen tavoitetaso, joka itsessään jää epämääräiseksi (”judged to be acceptable”), täyttyisi riittävällä (”appropriate”) tasolla (artikla 9). Mikäli järjestelmä kuuluisi EU:n harmonisoidun lainsäädännön alaan, jonka perusteella sen vaatimustenmukaisuuden eli myös riskienhallintatoimien asianmukaisuuden arvioinnin suorittaisi ulkopuolinen taho, arviointivastuu vaikuttaisi kohdentuvan instansseille, jotka osaavat parhaiten minimoida tuoteturvallisuuteen liittyviä riskejä[475]. Onkin mahdollista, että tekoälyasetuksen mahdollisuudet ihmisoikeusriskien tunnistamiseksi ja hallitsemiseksi jäävät toivottua heikommiksi.

Verrattain heikot keinot ihmisoikeusriskien hallitsemiseksi juontuvat sääntelyn perusratkaisusta. Asetuksessa tekoäly nähdään nimenomaan teknisenä tuotteena, ja sääntelyä rakennetaan tuoteturvallisuus-, ihmisoikeus-, kuluttajansuojasääntelyiden ominaispiirteitä yhdistellen[476]. Sillä siis pyritään sääntelemään tekoälytuotteiden markkinoita ja varmistamaan markkinoille saatettavien tuotteiden välittömien riskien arviointi ja hallinta tavalla, joka ei rajoita kilpailua tai markkinatoimintaa yli tarvittavan[477]. Suuririskisten sovellusten tunnistetusta riskipotentiaalista huolimatta säädöksessä keskitytään sääntelemään järjestelmien tuotantoprosessia [478] eikä lähtökohtaisesti vaadita, että lopullisen järjestelmän turvallisuuden tai lainmukaisuuden arvioisi riippumaton taho ennen markkinoille saattamista.

Riskien hallinnoimiseksi luotu, läpinäkyvyysvaatimusten varaan rakentuva protokolla perustuu vahvasti tekoälyjärjestelmien tarjoajien itsearviointeihin, minkä taustalla lienee pyrkimys markkinoiden mahdollisimman vähäiseen rajoittamiseen.

475 Castets-Renard, C., & Besse, P. (2023). Ex ante Accountability of the AI Act: Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance. *Pursuit of Fundamental Rights in the Country of Compliance. Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime. Comparative Law Perspectives. C. Castets-Renard & J. Eynard (eds),* Bruylant, Forthcoming.

476 Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 98–99.

477 Samaan tavoitteeseen liittyy myös direktiiviehdotus tekoälyyn liittyvästä vastuusta. Ks. Euroopan komissio. COM/2022/496 final. Proposal for AI Liability Directive.

478 Ks. metasääntelystä Viljanen, M. (2017). Algoritmien haaste: uuteen aineelliseen oikeuteen? *Lakimies 115 (2017): 7–8.* 1079–1080.

Niin säädöksen täytäntöönpanoon kuin säädösvelvoitteiden seuraamiseen liittyvät vaatimukset muodostuvatkin itsearvioinnin mahdollistamisen ansiosta kevyemmiksi kuin ne olisivat muodostuneet, mikäli ulkopuolisia arviointeja vaadittaisiin sääntelyn noudattamiseksi. Keveydellä on kääntöpuolensa. Artiklan 6 kohdan 2 (a) mukaan tekoälyjärjestelmän tarjoaja voi myös arvioida, että huolimatta siitä, että järjestelmä sinänsä kuuluu asetuksen perusteella korkean riskin kategoriaan, se ei aiheuta merkittävää riskiä haittojen syntymiselle (”significant risk of harm”). Tämä on mahdollista sellaisten järjestelmien kohdalla, jotka on tarkoitettu kapea-alaisten menettelytehtävien suorittamiseen (”narrow procedural task”), aiemmin suoritetun ihmisvetoisen toiminnan parantamiseen, tehtyjen päätösten kaavamaisuuden tai poikkeamien havaitsemiseen tai valmisteleviin tehtäviin. Jokainen kategoria on varsin epämääräinen, ja epämääräisyyttä lisää entisestään se, että järjestelmää arvioidaan sen perusteella, mihin se *on tarkoitettu* (”intended to”), eikä sen perusteella, mihin sitä voidaan käyttää. Mikäli tarjoaja arvioi järjestelmän kuuluvan johonkin näistä poikkeustilanteista, korkean riskin tekoälyjärjestelmien velvoitteet eivät kohdistu siihen. Ulkopuolinen taho ei tässäkään tilanteessa tarkista tarjoajan arvion asianmukaisuutta. Vaikka tarjoajan on rekisteröitävä tällainen järjestelmä ja varauduttava esittämään kansalliselle viranomaiselle arvio, johon päätös jättää järjestelmä korkean riskin tekoälyjärjestelmien velvoitteiden ulkopuolelle perustuu, säännönmukaisen valvonnan ollessa kansallisten viranomaisten oman aktiivisuuden varassa, riskit mahdollisille virhearvioinneille ja niistä seuraaville haitoille voivat muodostua merkittäviksi.

5.2.3 Yleiskäyttöisten tekoälymallien sääntelykehikko

Säädöksessä on oma osionsa yleiskäyttöisille tekoälymalleille, joita voidaan liittää monentyyppisiin järjestelmiin ja siten käyttää moniin erilaisiin tehtäviin. Monipuoliset käytön mahdollisuudet tekisivät tekoälyjärjestelmän käyttötarkoitukseen nojautuvan sääntelykehikon soveltamisen tällaisiin tekoälymalleihin jokseenkin mahdottomaksi: moneen taipuvan mallin tarjoajalla on varsin heikot mahdollisuudet arvioida, mihin mallia tullaan käyttämään ja siten myös sitä, mihin riskikategoriaan malli kuuluisi. Yleiskäyttöisiin malleihin katsottiin kuitenkin voivan liittyvän erityisiä, systeemisiä riskejä. Systeemiset riskit ovat makrotason riskejä, jotka artiklan 3(65) mukaan kohdistuvat kansanterveyteen, yleiseen turvallisuuteen, perusoikeuksiin tai yhteiskuntaan kokonaisuutena. Resitaalissa 110 tarkennetaan, että systeemiset riskit eivät kuitenkaan rajoitu nimenomaisesti mainittuihin.

Yleiskäyttöisten tekoälymallien sääntely puuttui komission alkuperäisestä ehdotuksesta. Vuonna 2021, kun komissio antoi ehdotuksen tekoälyasetukseksi, generatiivinen tekoäly ei ollut vielä noussut poliittisesti merkittäväksi kysymykseksi. Sen sääntelyn tarpeeseen herättiin toden teolla vasta sääntelytyön loppuvaiheessa vuoden 2023 aikana, kun erityisesti ChatGPT:n julkaiseminen nosti generatiivisen tekoälyn

erityispiirteistä kumpuavat kysymykset julkiseen keskusteluun. Kolmikantaneuvotteluissa jouduttiinkin luomaan yleiskäyttöisten tekoälymallien sääntelykehikko varsin kiireellä parlamentin ja neuvoston esittämien ehdotusten perusteella.

Yleiskäyttöisiä tekoälymalleja säädellään asetuksen luvussa V. Yleiskäyttöiset tekoälymallit jaetaan niiden kehitystyössä käytetyn laskentatehon perusteella "tavanomaisiin" yleiskäyttöisiin malleihin, joita koskevat tietyt läpinäkyvyysvaatimukset, sekä "systeemisesti riskialttiisiin" malleihin (art. 51), joita koskevat läpinäkyvyysvaatimusten lisäksi vaatimukset esimerkiksi mallin testaamisesta, systeemisten riskien arvioimisesta ja mahdollisten onnettomuuksien seuraamisesta. Malli katsotaan systeemisesti riskialttiiksi silloin, kun sen kehittämisessä käytetty laskentateho ylittää määritellyt, hyvin korkeaksi säädetyt raja-arvot[479]. Ilmeisesti vain "kehittyneimmät" tekoälysovellukset hyvin tiukasti ajatellen haluttiin systeemisesti riskialttiiden mallien sääntelyn piiriin. Muiden kuin systeemisesti riskialttiiden yleiskäyttöisten tekoälymallien kohdalla avoin ja muokattavissa oleva lähdekoodi keventää säädösvelvoitteita.

Läpinäkyvyysvaatimukset, jotka koskevat yleiskäyttöisen tekoälymallin tarjoajia riippumatta siitä, onko yleiskäyttöinen tekoälymalli systeemisesti riskialtis vai ei, pitävät sisällään vaatimukset varsin laajasta teknisestä dokumentaatiosta ja suppeammasta kehittäjäinformaatiosta. Teknisen dokumentaation on sisällettävä liitteen XI mukaiset tekniset tiedot mallin kehitystyöstä ja ominaisuuksista. Tekninen dokumentaatio on pyydettäessä esitettävä kansalliselle valvontaviranomaiselle tai tekoälyviranomaiselle. Kehittäjäinformaation vaadittu sisältö kuvataan liitteessä XII, ja se jää vaatimattomaksi tekniseen dokumentaatioon verrattuna. Mallin tarjoajan on esitettävä se tahoille, jotka haluavat liittää mallin osaksi tekoälyjärjestelmää, ja siihen tulee sisällyttää tiedot, joita näiden on arvioitu tarvitsevan säädösvelvoitteista suoriutumiseksi. Lisäksi mallien tarjoajien on julkaistava riittävän yksityiskohtainen ("sufficiently detailed") tiivistelmä mallin opetukseen käytetystä datasta. Tiivistelmän tulee olla julkisesti saatavilla. Näiden velvoitteiden lisäksi säädöksessä nostetaan erikseen esille, että yleiskäyttöisten tekoälymallien tuottajien on toimittava EU:n tekijänoikeuslainsäädännön mukaisesti.

Systeemisesti riskialttiiden mallien tarjoajien on kaikkia yleiskäyttöisiä tekoälymalleja koskevien velvoitteiden lisäksi arvioitava malliin liittyviä systeemisiä riskejä, testattava malli niin sanotun "red-team" -testaamisen eli vihamielisen testaamisen keinoin ja dokumentoitava testaaminen. Tarkoituksena on tunnistaa malliin mahdollisesti liittyvät systeemiset riskit, joita tarjoajan on pyrittävä vähentämään. Tarjoajan on myös seurattava mahdollisia onnettomuuksia ja huolehdittava kyberturvallisuudesta ja fyysisen infran riittävyydestä.

[479] Tekninen raja päätettiin asettaa 10^25 FLOPiin (floating points operations).

Systeemisesti riskialttiiden yleiskäyttöisten tekoälymallien sääntely vaikuttaa siis ensisilmäyksellä tiukalta. Se on kuitenkin keventynyt parlamentin esittämästä. Parlamentti olisi vaatinut, että yleiskäyttöisten mallien tarjoaja hyödyntäisi mallin arvioinnissa "riittäviä menetelmiä" ("appropriate methods"), joihin se olisi lukenut esimerkiksi riippumattomien asiantuntijoiden avun. Tällaista vaatimusta lopullisessa versiossa ei ole, vaan arviointi voidaan toteuttaa sisäisesti. Systeemisesti riskialttiiden tekoälymallien tarjoajien on tunnistettava systeemiset riskit ja vähennettävä niitä, mutta riskien hyväksyttävää tasoa ei määritellä. Lisäksi systeemisten riskien arviointivelvollisuus kohdennetaan hyvin kapea-alaisesti. Teknisten rajojen perusteella ainoastaan kehittyneimmät tekoälymallit yltävät systeemisesti riskialttiiden mallien kategoriaan, EU-komission aiemman tiedotuksen mukaan tällä hetkellä ainoastaan Googlen Gemini ja GPT4[480]. Huomionarvoista on myös se, että vaikka systeemisesti riskialttiisiin malleihin tunnistetaan liittyvän vakavia riskejä, näistä riskeistä huolimatta tekoäly*järjestelmiä,* joihin tällainen malli liitetään, ei välttämättä arvioida suuririskisiksi, sillä riskitaso määräytyy järjestelmän käyttötarkoituksen, ei siinä hyödynnettävien tekoälymallien perusteella.

5.2.4 Sääntelyn arviointia

Sääntelyä heikentävät yhtäältä massiiviset rajaukset sääntelyn soveltamisalassa ja toisaalta poikkeukset ja liennytykset, joita sinänsä kiellettyihin tekoälysovelluksiin on liitetty. Erityisesti rajaus, jonka puitteissa kaikki kansalliseen turvallisuuteen ja maanpuolustukseen liittyvät toiminnot on rajattu kokonaan sääntelyn ulkopuolelle, hapertaa sääntelyn vaikuttavuutta. Toisaalta kansallinen turvallisuus on jo SEU (sopimus Euroopan unionista) 4 artiklassa määrätty nimenomaan jäsenvaltioiden oman säädösvallan alaiseksi alueeksi, johon EU-sääntelyllä ei tule puuttua. Euroopan ihmisoikeustuomioistuin on kuitenkin suhtautunut jokseenkin pidättyväisesti kansalliseen turvallisuuteen vetoamiseen[481] – esimerkiksi jo vuonna 2013 EU:n ihmisoikeustuomioistuimen tutkimusryhmä kuvasi, että ihmisoikeustuomioistuimen käytännössä oli jo pitkään vaadittu, että kansalliseen turvallisuuteen vetoava valtio pystyy osoittamaan, että toiminta tosiasiallisesti liittyy kansalliseen turvallisuuteen eikä perusoikeuksia vähempää rajoittavia vaihtoehtoja ole[482]. Vuonna 2020 tapauksessa La

480 Komission tiedotuksesta on poistunut viittaukset olemassaoleviin, raja-arvot täyttäviin malleihin, mutta esimerkiksi Rochierin uutisista on vielä löydettävissä viittaus komission mainintaan. Ks. Rochier. (2024). The European Parliament adopts the Artificial Intelligence Act. (29.4.2024). *Rochier Insights.*

481 Euroopan ihmisoikeustuomioistuimen tutkimusosasto. (2013). *National security and European case-law.* Euroopan neuvosto/ECHR.

482 Ibid. 40.

Quadrature du Net[483] EU-tuomioistuin katsoi, että ”pelkästään se, että kansallinen toimenpide on annettu kansallisen turvallisuuden suojaamiseksi, ei voi johtaa siihen, ettei unionin oikeutta sovelleta”[484], ja erotti yleiseen turvallisuuteen kohdistuvat eli kansalliseen turvallisuuteen liittyvät rikokset, kuten terrorismin, vakavista rikoksista, jotka vakavuudestaan huolimatta eivät horjuta kansallista turvallisuutta.

EU:n oikeuskäytännössä omaksuttu linja kuitenkin koskee nimenomaisesti SEU 4 artiklasta esiin nousevaa rajausta, jota sovellettaisiin myös tekoälyasetuksen alueella ilman asetustekstiin liitettyä erillistä rajoitusta. On mahdollista, että tekoälyasetuksessa omaksuttu rajaus mahdollistaa laajemman vetoamisen kansalliseen turvallisuuteen – erityisesti, kun asetustekstissä painotetaan, että rajaus koskee tekoälyjärjestelmien markkinoille saattamista, käyttöönottamista ja käyttämistä yksinomaan sotilaallisiin, puolustuksellisiin tai kansallisen turvallisuuden tarkoituksiin, *riippumatta siitä, mikä taho näin tekee*. Käytännössä asetusteksti vapauttaneekin säädösvelvoitteista tekoälyjärjestelmien tarjoajat, jotka tuovat markkinoille asetuksen perusteella suuririskisiä tai kiellettyjä järjestelmiä, jotka on kehitetty kansallisen turvallisuuden varmistamiseen. Kun tällaisten järjestelmien kehityksessä ei tarvitse noudattaa tekoälyasetuksen vaatimuksia, riskinä on, että kansallisen turvallisuuden varmistamiseksi tuotetaan heikkolaatuisia, mahdollisesti myös turvallisuuden kannalta kyseenalaisia järjestelmiä[485]. Rajauksen todellinen merkitys kuitenkin selviää vasta oikeuskäytännössä, mitä voi pitää lainsäädännön selkeyden kannalta ongelmallisena. Ongelmallisuutta lisää entisestään se, että tekoälyjärjestelmän käyttökohteen määrittely ennakolta ei välttämättä ole yksioikoista: esimerkiksi biometriseen tunnistamiseen tai massavalvontaan kehitetty järjestelmä voi olla käytettävissä paitsi terrorismin ehkäisemiseen eli selkeästi kansallisen turvallisuuden alaan kuuluvaan tehtävään, myös moniin muihin käyttötarkoituksiin. Käytännön tasolla voikin olla hyvin vaikea määritellä, milloin tekoälyjärjestelmä on tarkoitettu ”exclusively for” sotilaallisiin, maanpuolustuksellisiin tai kansalliseen turvallisuuteen liittyviin käyttötarkoituksiin.

Asetuksen riskiperustainen lähestymistapa, joka keskittyy riskien tunnistamiseen ja ennakolliseen hallintaan, jättää hyvin vähälle huomiolle odottamattomat seuraukset. Se luo illuusion siitä, että algoritmisten järjestelmien vaikutukset voisivat olla kattavalla tasolla ennakoitavissa[486]. Kiihtyvällä vauhdilla algoritmisoituvassa ja

483 EU-tuomioistuin, tapaus C-511/18: La Quadrature du Net.

484 Ibid. Kappale 99.

485 Ks. myös Day, J., Iwańska, K., Simon, E. & Willamo, K. (2024). *Packed with loopholes: Why the AI act fails to protect civic space and the rule of law.* Civil Liberties Union for Europe e.V.

486 Lavorgna, A., & Ugwudike, P. (2022). Managing risks, passing over harms? A commentary on the proposed EU AI Regulation in the context of criminal justice. *Justice, Power and Resistance*, *5*(3), 292–298.

yhä nopeammin muuttuvassa yhteiskunnassa riskien arvioiminen käynee kuitenkin jatkuvasti vaikeammaksi ja arvioinnin ajallinen relevanssi lyhyemmäksi. Vaikka sääntelyssä tunnistetaan, että tekoälyyn liittyy tietyissä tapauksissa riski systeemisille haitoille, niiden rajoittamiseksi vaaditut toimet jäävät lähes olemattomiksi ja kohdistuvat äärimmäisen harvoihin tekoälymalleihin. Tekoälysäädös rajautuukin melko kapea-alaisesti hallinnoimaan sellaisia riskejä, joiden kohde ja aiheuttaja ovat helposti tunnistettavissa ja määriteltävissä. Yleiskäyttöisiä tekoälymalleja sääntelevässä luvussa V sen sijaan vaaditaan systeemisesti riskialttiiden yleiskäyttöisten tekoälymallien tarjoajia arvioimaan laaja-alaisesti myös makrotason riskejä. Tällaisten riskien hallintaan ei säädös kuitenkaan tarjoa työkaluja. Makrotason riskien olemassaolon tunnustaminen ja tunnistaminen on tervetullut parannus alkuperäisestä komission ehdotuksesta, mutta velvoite arvioida ja vähentää riskejä jää melko vaatimattomaksi, kun säädöksessä ei määritellä sallitun tai kielletyn riskitason rajoja. Lisäksi riskien tunnistamisen ja vähentämisen vaatimukset rajataan kohtuudella ennakoitavissa oleviin (”reasonably foreseeable”). Epäselväksi jää, kuinka tiukka määritelmä on, ja missä tilanteissa esimerkiksi epäsuorat riskit voitaisiin arvioida kohtuudella ennakoitavissa oleviksi.

Sääntelyssä ei myöskään onnistuta kehittämään työkaluja, joiden avulla pystyttäisiin hillitsemään tekoälyteknologioihin liittyvää kiihtyvää *muutosta*. Esimerkiksi Hasselbalch on nostanut muutoksen hallinnan ongelmat esille ja argumentoinut, että teknologian sääntelemisen keinot, jotka painottuvat teknologioiden arviointimenettelyihin *ex ante* ja/tai vaikutusten arviointiin *ex post,* eivät tunnista tai kykene hallitsemaan teknologista kehitystä, joka tapahtuu vähittäin ajassa, ei äkillisesti[487]. Jotta teknologioiden vaikutuksia voitaisiin hallita, sääntelyn keinoin tulisi voida vaikuttaa myös teknologisen kehityksen suuntaan sen sijaan, että keskityttäisiin arvioimaan ainoastaan yksittäisten teknologioiden ominaisuuksia ja vaikutuksia irrallaan teknologisen kehityksen suuremmista linjoista.

Tekoälysäädöksessä painotetaan turvallisuutta ja pyritään puuttumaan ihmisoikeusriskeihin, ja suojelun kohteeksi nostetaan nimenomaan yksilöt ja ryhmät – joille kuitenkin annetaan hyvin rajatusti mahdollisuuksia hakea ja saada oikeutta[488]. Vaikka yksilöt voivat esittää valituksia tekoälyjärjestelmistä, niiden vaikuttavuus jäänee vähäiseksi, sillä viranomaisten on ainoastaan huomioitava valitukset markkinavalvontatoimissaan, ei esimerkiksi vastattava niihin (art. 85).

487 Hasselbalch, J. A. (2018). Innovation assessment: governing through periods of disruptive technological change. *Journal of European Public Policy*, *25*(12), 1855–1873.

488 Ks. myös Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft U Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 111.

Epäsuorien haitallisten vaikutusten hallinta lainsäädännön keinoin voi olla erittäin vaikea tehtävä. Kun tekoälyjärjestelmät tuovat mukanaan riskin valtavan nopealle, ennakoimattomalle ja kauaskantoiselle muutokselle, tehokas sääntely on kuitenkin välttämätöntä. Ilman tehokasta sääntelyä kokonaisvaltaisesti ihmisiin ja yhteiskuntaan vaikuttavien teknologioiden hallitsematon leviäminen ja hyödyntäminen voi vaarantaa yhteiskuntien vakauden, oikeusvaltioperiaatteen ja demokratian toteutumisen[489]. Sen sijaan, että demokraattinen valtio hallitsisi yhteiskunnan kehitystä ja toimintaa, valta siirtyisi tekoälyjärjestelmille ja niiden tuottajille. Osaltaan niin voi katsoa jo käyneen. EU-sääntelyllä olisi mahdollisuus turvata yhteiskuntien toimintamahdollisuudet, mikäli se pystyisi tehokkaasti huomioimaan tekoälyjärjestelmiin liittyvät, myös epäsuorat riskit ja haitat, joihin tekoälyn laaja kehitys ja käyttö vaikuttaisivat johtavan. Lopullisessa sääntelyssä tekoälyjärjestelmien potentiaaliset ja jo todistetut haitalliset vaikutukset jäävät kuitenkin heikosti tunnistetuiksi ja niihin puuttuminen vaillinaiseksi.

489 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, 10.

6 Lopuksi

6.1 Yhteenveto

Tässä johdanto-osassa tavoitteenani on ollut arvioida ensinnäkin, minkälaisia yhteiskunnallisia haittoja algoritmisiin teknologioihin liittyy, toiseksi, miten algoritmisten teknologioiden ominaisuudet vaikuttavat haittoihin, ja kolmanneksi, miten sääntelyssä tunnistetaan haitat ja pyritään hallitsemaan niitä.

Ensimmäisessä luvussa pohjustin tutkimusasetelmaani ja avasin teoreettista viitekehystä, joka on ohjannut tutkimukseni toteuttamista. Lainopillisesta painotuksesta huolimatta tutkimukseni teoreettinen pohjavire on yhteiskuntatieteellinen ja se asettuu osaksi yhteiskunnallisten haittojen tutkimusperinnettä. Algoritmisten haittojen yhteiskunnallisen ulottuvuuden kattavammaksi ymmärtämiseksi olen soveltanut Rosan yhteiskunnallisen kiihtyvyyden teoriaa[490] ja Giddensin rakenteistumisteoriaa[491]. Niiden kautta yhteiskunta näyttäytyy paitsi yksilöiden toimintaa määrittävänä rakenteena, myös yksilöiden toiminnan aikaansaamien muutosten kohteena. Tällöin yhteiskuntaa on mahdollista tarkastella myös haittojen kohteena.

Toisessa luvussa tiivistin väitöskirjan kolme osatutkimusta. Ensimmäinen osatutkimus[492] osoittaa, että algoritmiset teknologiat muuttavat yhteiskunnallisten haittojen ilmenemistä niin laadullisesti kuin määrällisesti. Toisessa osatutkimuksessa[493] esitetään, että lainsäädäntötyössä korostuvat diskurssit, joissa tekoälyn avaamia mahdollisuuksia painotetaan, kun taas tekoälyyn liittyviä riskejä tai haittoja tarkastellaan ainoastaan rajoitetusti. Tämä vaikuttaisi linkittyvän jokseenkin vahvasti yhteiskunnan kiihtyvään vauhtiin ja siitä seuraavaan tarpeeseen tehostaa hitaasti toimi-

490 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

491 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

492 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

493 Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.

via byrokraattisia prosesseja. Kolmannessa osatutkimuksessa[494] huomio kiinnitetään yhtäältä tekoälypolitiikan idealististen tavoitteiden ja lainsäädännössä asetettujen velvoitteiden välisiin ristiriitoihin, toisaalta poliittisten sektoreiden siiloutumiseen, joka näyttää johtaneen kokonaiskuvan hahmottamisen ongelmiin.

Kolmannessa luvussa arvioin tekoälyjärjestelmiin liittyviä yhteiskunnallisia haittoja. Haittojen luokittelemiseksi yhdistelin erilaisia haittojen arviointiin kehitettyjä viitekehyksiä. Yhteiskunnallisten haittojen teoria auttoi tunnistamaan haittojen ilmenemismuotoja[495]. Teorian tukena soveltamani teknologisten haittojen viitekehys[496] auttoi luokittelemaan algoritmiset haitat välittömiksi ja seurannaishaitoiksi. Välittömät haitat määrittelin haitoiksi, jotka seuraavat välittömästi teknologian käytöstä, kun taas seurannaishaitat määrittelin haitoiksi, jotka syntyvät teknologian välittömän vaikutuspiirin ulkopuolella, mutta joiden syntymisessä teknologian olemassaolo ja toiminta on väistämätön edellytys.

Kolmannen luvun perusteella on nähtävissä, että aiemmassa tutkimuksessa algoritmisista haitoista on tunnistettu parhaiten yksilöihin ja ryhmiin kohdistuvat välittömät haitat, joiden syntymisessä algoritmisen teknologian toiminnan tavat ovat määräävässä asemassa. Algoritmisista teknologioista juontuvat seurannaishaitat, joiden syntymisessä korostuvat algoritmisen järjestelmän toimintatapojen lisäksi sen käyttämisen tavat ja käyttökonteksti, ovat vielä heikosti tunnettuja ja vasta vähitellen saamassa ansaitsemaansa huomiota. Lisäksi näyttää siltä, että algoritmiset haitat, jotka kohdistuvat yhteiskuntaan ja siis heikentävät yhteiskunnan mahdollisuuksia turvata jäsenilleen hyvän elämän edellytyksiä, eivät ole vielä juurikaan nousseet tutkimuksen kohteeksi oikeustieteen, kriittisen kriminologian tai haittojen tutkimuksen kentällä. Tutkimus tekoälyteknologioiden aiheuttamista, yhteiskuntaan vaikuttavista haitoista painottuu sen sijaan esimerkiksi filosofian ja sosiologian aloille.

Neljännessä luvussa keskiössä olivat tekoälyteknologioiden tekniset ominaisuudet ja niiden vaikutukset haittojen muodostumiseen Viljasen[497] näkemysten inspi-

494 Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3), 270–291.

495 Tombs, S. ja Hillyard, P. (2004). Towards a political economy of harm: States, corporations and the production of inequality. Teoksessa Hillyard, P., Pantazis, C., Tombs, S. & Gordon, D. (eds). *Beyond Criminology: Taking harm seriously*: 30–54. London: Pluto Press;

Pemberton, S. (2015). Harmful Societies: Understanding Social Harm. Policy Press.

496 Wood, M. A. (2021). Rethinking how technologies harm. The British Journal of Criminology, 61(3), 627–647.

497 Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.

roimana. On selvää, että tekoälyjärjestelmien ominaisuudet vaikuttavat tekoälyjärjestelmien haittapotentiaaliin. Etenkin välittömiä haittoja syntyy tekoälyteknologioiden virheellisen, vinoutuneen tai odottamattoman toiminnan seurauksena. Riski haittojen syntymiselle on luonnollisesti sitä suurempi, mitä vaikeampi ihmisen on yhtäältä ymmärtää, ennakoida ja kontrolloida järjestelmän toimintaa tai toisaalta varmistua riittävällä tasolla sen asianmukaisuudesta esimerkiksi testaamisen avulla. Näihin mahdollisuuksiin taas vaikuttavat teknologian ominaisuudet: ainakin Viljasen hahmottelemat itsenäisyys, monimutkaisuus, selittämättömyys, epälineaarisuus, indeterministisyys ja dynaamisuus[498], sekä monikäyttöisyys. Koska välittömät haitat juontuvat verrattain usein teknologian toiminnan epätarkoituksenmukaisuudesta, vinoumista tai odottamattomuudesta, niiden voi usein nähdä kiinnittyvän teknologian suunnitteluaikaisiin valintoihin sekä niistä juontuviin järjestelmän ominaisuuksiin. Teknologian ominaisuudet taas näyttäisivät vaikuttavan jokseenkin suoraviivaisesti välittömien haittojen ilmenemisen riskiin.

Seurannaishaittojen syntymisessä merkityksellisten tekijöiden joukko kasvaa. Samoin kasvaa myös teknologian käytön ja siitä mahdollisesti seuraavan haitan etäisyys. Seurannaishaittoja voi syntyä, vaikka algoritminen teknologia toimisi asianmukaisesti ja suunnitellusti eikä sen toiminta aiheuttaisi välittömiä haittoja. Tällöin teknologian tekniset ominaisuudet tai sinänsä virheetön ja kontrolloitu toiminta eivät välttämättä juurikaan vaikuta haittapotentiaalin suuruuteen. Merkitystä on sen sijaan erityisesti sillä, millaisessa kontekstissa ja miten tekoälyteknologioita käytetään. Nämä seikat vaikuttavat siihen, minkälaisia vaikutuksia teknologian käyttämisellä on sen ympäristöön ja käyttäjiin. Teknologian tekniset ominaisuudet määrittelevät rajat sille, miten teknologiaa on mahdollista käyttää, mutta ne eivät itsessään riitä selittämään seurannaishaittojen syntyä.

Viidennessä luvussa käsiteltiin tekoälyyn liittyvää lainsäädäntöä. Luvun keskeisenä tavoitteena oli arvioida, miten lainsäädännössä tunnistetaan algoritmisten haittojen potentiaali ja miten olemassa olevat lait antavat mahdollisuuksia estää tai vähentää haittojen ilmenemistä.

Luvun ensimmäinen alaluku keskittyi kansalliseen lainsäädäntöön. Suomessa hallinnollisessa päätöksenteossa sallittu automaatio on rajattu niin tiukasti, että EU:n tekoälyasetuksen määritelmän mukaan se ei lukeudu tekoälyksi ylipäätään. Vaikuttaakin siltä, että etenkin koneoppivien algoritmisten järjestelmien riskipotentiaali on kansallisesti tunnistettu: hallintolain luvussa 8b sallitaan ainoastaan sääntöpohjainen automaattinen päätöksenteko, jossa järjestelmään ohjelmoidut säännöt perustuvat lakiin tai vakiintuneisiin käytäntöihin. Koneoppimiseen pohjautuvien automaattisten päätöksentekojärjestelmien käyttäminen on kielletty. Lisäksi lainsäädännössä on

498 Ibid.

määritelty viranomaisille velvoitteita, joiden tarkoituksena on varmistaa automaation asianmukainen toteuttaminen, sekä annettu automaattisen päätöksenteon kohteelle mahdollisuus maksuttomaan oikaisuvaatimukseen.

Kansallinen sääntely näyttää jokseenkin onnistuneelta automaattisten päätöksentekojärjestelmien haittapotentiaalin rajaamisen kannalta katsottuna. Toisaalta siinä on haittojen hallitsemisen kannalta myös joitakin selkeitä ongelmia. Ensinnäkin hallintolaki mahdollistaa rajoituksetta kaikenlaisten tekoälyjärjestelmien käytön hallinnollisen päätöksenteon tukena niin kauan kuin lopullinen päätös on ihmisvirkailijan tekemä. Käytännössä sääntely mahdollistaa siis esimerkiksi päätösten luonnostelun tekoälytyökalujen avulla. Näin ollen tekoälyjärjestelmät voivat vaikuttaa merkittävällä tavalla viranomaisten käytänteiden ja päätöksenteon taustalla ilman, että päätösautomaatioon lainsäädännössä liitetyt velvoitteet realisoituvat. Riski erityisesti epäsuorille algoritmisille haitoille on ilmeinen. Toiseksi, kun viranomaisille on annettu mahdollisuus määritellä vakiintuneiden käytäntöjen perusteella tyyppitapauksia, jotka voidaan automatisoida, on mahdollista, että hallinnollisessa päätöksenteossa siirrytään joissakin tilanteissa lähelle niin sanottua massaharkintaa[499], vaikka hallintopäätösten tulee hallintolain mukaan perustua yksilölliseen harkintaan.

Toisessa alaluvussa käsittelin EU:n tekoälyasetusta. Kun EU-lainsäädäntö tulee sovellettavaksi, kansallinen lainsäädäntö luonnollisesti väistyy sen tieltä. Asetuksen 3(1) artiklassa tekoälyksi määritellään vähintään jossain määrin itsenäisesti toimivat konepohjaiset järjestelmät, jotka päättelymoottorin avulla pystyvät tuottamaan esimerkiksi suosituksia, ennusteita tai muunlaista sisältöä, ja siten voivat vaikuttaa ympäristöihin, joiden kanssa ne ovat vuorovaikutuksessa. Sääntelyn avulla pyritään hallitsemaan tekoälyjärjestelmistä kumpuavia riskejä sääntelemällä tekoälysovellusten tuotantoa, markkinoille saattamista ja käyttöä. Asetus on riskiperusteinen, ja tekoälyjärjestelmät luokitellaan siinä neljään riskiluokkaan: kielletyn, suuren, rajatun ja pienen riskin järjestelmiksi. Lisäksi yleiskäyttöisille tekoälymalleille on luotu oma riskiluokituksensa, jonka lähtökohtana on, että kehittyneimmät yleiskäyttöiset tekoälymallit tuovat mukanaan systeemisiä riskejä. Riskitason määrittämisestä vastaa asetuksen mukaan järjestelmän tai mallin tuottaja. Suuren riskin järjestelmien tuotantoa säädellään melko tiukasti, sitä vähäisemmän riskin sovelluksia taas jokseenkin kevyesti. Yleiskäyttöisten tekoälymallien sääntely nojaa pääasiallisesti läpinäkyvyysvelvoitteisiin.

Välittömien haittojen hallitsemisen kannalta sääntelystä voi muodostua melko toimiva kokonaisuus. Suuririskisiksi määriteltyjen sovellusten tuotannon sääntelyllä pystytään todennäköisesti lisäämään tällaisten järjestelmien toiminnan ymmärrettävyyttä, ennakoitavuutta ja kontrolloitavuutta, mikä pienentänee tekoälyjärjestelmien

499 Koivisto, I., & Koulu, R. (2020). Miten hyvä hallinto digitalisoidaan? Haaste oikeustieteelliselle tutkimukselle. *Lakimies*, *118*(6), 798–821.

ominaisuuksista kumpuavaa potentiaalia välittömien haittojen aiheutumiselle. Sääntely vähentänee myös riskiä tuotannonaikaisille virheille ja puutteille sekä niistä seuraaville haitoille. Toisaalta, kun sääntelyssä keskitytään kontrolloimaan ennakoitavia, ihmisoikeuksiin kiinnittyviä *riskejä*, käsitys *haitoista* rajautuu melko teknisen ihmisoikeuskehyksen mukaisesti[500]. Tämän takia sääntelyn tarjoamat mahdollisuudet tunnistaa ja vähentää haittoja, jotka syntyvät muutoin kuin nimenomaisten oikeudenloukkauksien seurauksena, jäävät vähäisiksi. Lisäksi monet ehdotetut velvoitteet muodostuvat epämääräisiksi, ja tuottajille annetaan laaja valta arvioida, minkälaiset toimet ovat sopivia ja riittäviä riskien vähentämiseksi.

Seurannaishaittojen hallitsemiseen tekoälyasetus tarjoaa vain vähän työkaluja. Asetuksessa keskitytään varmistamaan, että etenkin suuren riskin sovellusten tuotantoprosessi on tarkasti hallittu ja tekoälyteknologiat laadukkaasti tuotettu. Riskien hallintaa pyritään siis toteuttamaan suurelta osin *ex ante* – vaikka lienee jokseenkin selvää, että ennakollisen riskiarvioinnin keinoin ainoastaan osa potentiaalisesti syntyvistä haitoista pystytään tunnistamaan. Erityisesti seurannaishaittojen muodostumisessa järjestelmän teknisillä ominaisuuksilla tai tuotannon tavoilla ei ole välttämättä suurta merkitystä: seurannaishaittoja voi syntyä myös hyödyllisen, asianmukaisesti tuotetun ja tarkoituksenmukaisesti toimivan sovelluksen epäsuorana seurauksena. Tällaiset epäsuorat seurannaishaitat jäävät sääntelyssä tunnistamatta. Tekoälyasetus on laaja ja vaikeasti hallittava, ja lienee todennäköistä, että vasta vuosien kuluttua, mahdollisesti lukuisten EU-tuomioistuimen kannanottojen jälkeen, selviää sääntelyn todellinen merkitys algoritmisten haittojen hallinnassa.

6.2 Keskustelu

6.2.1 Yhteiskunnallisten haittojen teoria ja algoritmiset teknologiat

Perinteisesti yhteiskunnallisten haittojen tutkimuksessa ovat korostuneet välittömät, yksilöiden kohtaamat haitat, jotka tuotetaan yhteiskunnallisissa prosesseissa. Tällaiset haitat, myös silloin kun ne seuraavat algoritmisten teknologioiden käytöstä, näyttävät olevan melko laajasti tunnistettuja ja tutkittuja, joskin yhteiskunnallisten haittojen tutkimusperinteessä algoritmisia haittoja on tutkittu vielä verrattain vähän. Tekoälyn haitallisia vaikutuksia arvioitaessa keskittyminen välittömiin haittoihin ei kuitenkaan riitä. Valtava muutosvoima, joka tekoälyyn ja algoritmisiin teknologioi-

[500] Vasta parlamentin kannassa haitta määriteltiin, ja määritelmä löytyy lopullisen asetuksen resitaalista 5. Haitan mittaamisen tapoja ei avata. Resitaalissa esitetään, että haitta voi olla aineellista tai aineetonta, ja fyysistä, psyykkistä, yhteiskunnallista tai taloudellista.

hin liittyy, muovaa ihmisiä ja yhteiskuntaa ennennäkemättömällä tavalla[501]. Seurauksena ilmenee myös yllättäviä ja odottamattomia vaikutuksia, joista osa on haitallisia. Seurannaishaitat kiinnittyvät algoritmisten teknologioiden transformatiivisiin ominaisuuksiin, ja ne voivat ilmetä ajallisesti ja maantieteellisesti hyvinkin kaukana teknologian välittömästä vaikutuspiiristä.

Vaikka myös algoritmisia seurannaishaittoja on jonkin verran tutkittu, kokonaiskuva algoritmisista haitoista on vielä epämääräinen. Teknologian välittömästä käyttöympäristöstä etääntyvät haitat ovat jääneet haittojen tutkimuksessa heikosti tunnistetuiksi ja tutkituiksi. Algoritmisten teknologioiden aiheuttamiin *muutoksiin* kytkeytyvät haitat voidaan kuitenkin saattaa haittojen tutkimuksen perinteeseen, jos heijastusvaikutusten[502] eli ensisijaisen haitan takia syntyvien uusien haittojen määritelmää laajennetaan siten, että tarkastellaan myös haittoja, joita syntyy tai voi syntyä teknologian sinänsä neutraalien tai positiivisten vaikutusten takia. Tällöin heijastusvaikutukset voidaan nähdä osana seurannaishaittoja, joiden synty kiinnittyy yhteiskunnallisen todellisuuden muutokseen. Algoritmisten seurannaishaittojen kohdalla arvioitavaksi siis tulevat haitat, joita syntyy, kun algoritmiset teknologiat vaikuttavat ympäristöön, jossa niitä käytetään, käyttäjiinsä tai tahoihin, joihin käytön vaikutukset kohdistuvat. Tällaisten haittojen syntymekanismeja on analysoinut Wood, jonka viitekehyksessä teknologian vaikutuksista seuraavia haittoja nimitetään generatiivisiksi haitoiksi[503].

Toisin kuin Wood, joka arvioi teknologian vaikutusmekaniikkaa haitan taustalla, määritelmäni seurannaishaitoista seuraa samaa logiikkaa kuin yhteiskunnallisten haittojen tutkimusperinteestä tuttu heijastusvaikutusten[504] määritelmä: kuten heijastusvaikutusten kohdalla, algoritmisten seurannaishaittojen tunnistamiseksi arvioidaan, mitä haittoja algoritmisen järjestelmän käytöstä tai toiminnasta voi seurata tai on seurannut *välittömien vaikutusten jälkeen*. Painotus on siis algoritmisen teknologian ja syntyneen haitan välisessä etäisyydessä. Toisin kuin haittojen tutkimusperinteen heijastusvaikutusten kohdalla, seurannaishaittoihin johtavien, algoritmisten teknologioiden aikaansaamien välittömien vaikutusten ei kuitenkaan tarvitse olla haitallisia.

Toisin kuin yksilöihin tai ryhmiin kohdistuvia, etenkin välittömiä haittoja, algoritmisista teknologioista johtuvia, yhteiskuntaa haitallisilla tavoilla muuttavia kehi-

501 Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

502 Tombs, S. (2019). Grenfell: the unfolding dimensions of social harm. *Justice, Power and Resistance 3/1*. 61–88.

503 Wood, M. A. (2021). Rethinking how technologies harm. The British Journal of Criminology, 61(3), 627–647.

504 Tombs, S. (2019). Grenfell: the unfolding dimensions of social harm. *Justice, Power and Resistance 3/1*. 61–88.

tyskulkuja on vasta alettu tunnistaa. Tällaisten kehityskulkujen yhteiskunnallinen analyysi on viime vuosina lisääntynyt. Yhteiskunnallisten haittojen tutkimusperinteessä yhteiskuntaa ei kuitenkaan ole vielä juurikaan arvioitu haittojen kohteena. Rajaukselle on sinänsä olemassa hyvät perusteet. Yhteiskunnalliset haitat viittaavat *yhteiskunnallisesti tuotettuihin* haittoihin, ja yhteiskunnallisten haittojen tutkijat ovat perinteisesti nähneet haittojen kumpuavan kapitalistisesta talousjärjestelmästä ja siihen sidotuista yhteiskunnallisista prosesseista[505]. Haitat sen sijaan näkyvät yhteiskunnan sisällä ja ihmisten elämissä.

Yhteiskunta voidaan kuitenkin asemoida haittojen tutkimuksen kentällä myös toisin. Yhteiskunnallisen kiihtyvyyden teorian[506] keskeinen argumentti on, että kapitalistinen talousjärjestelmä on ainoastaan yksi modernin yhteiskunnan kehityskulkuja ja toimintoja ohjaavista näkymättömistä voimista, joiden seurauksena haittoja yhteiskunnissa syntyy. Tämän lisäksi on huomioitava muut voimat: kulttuurinen paine tehostaa, nopeuttaa ja kiihdyttää arkielämää, sekä yhteiskunnan rakenteelliset tekijät, jotka ohjaavat toimintojen eriytymiseen. Etenkin kulttuurisen, kiihtyvyyttä ylläpitävän ja eskaloivan moottorin kautta yhteiskunnan jäsenet voidaan hahmottaa toimijoina, joiden valinnat tosiasiallisesti vaikuttavat yhteiskunnan kehityskulkuihin – jolloin toimijoilla on myös kyky *muuttaa* yhteiskuntaa, tarkoituksella tai tahatta, mahdollisesti myös entistä haitallisempaan suuntaan. Tämä lähenee Giddensin rakenteistumisteoreettista argumentaatiota, jossa yksilön ja yhteiskunnan suhde nähdään *duaalisena,* kahdensuuntaisena[507]. Kahdensuuntainen vuorovaikutus yhteiskunnan jäsenten ja yhteiskunnan rakenteellisten ominaisuuksien välillä on aiemmin jäänyt yhteiskunnallisten haittojen tutkimuksessa vähälle huomiolle. Mahdollisesti tämän takia yhteiskunnallisten haittojen tutkijat eivät ole juurikaan keskittyneet kysymykseen siitä, minkälaisia haittoja yhteiskuntaan kohdistuu.

Jos katsotaan, että yhteiskunta voi inhimillisen toiminnan seurauksena muuttua, on perusteltua tunnustaa myös, että aiheutetut tai aiheutuneet muutokset voivat olla haitallisia. Tällöin yhteiskunta hahmottuu paitsi haittoja tuottavana entiteettinä, myös potentiaalisena haittojen kohteena. Kun algoritmiset haitat kohdistuvat yhteiskuntaan, ne muuttavat yhteiskunnan rakenteita tai käytänteitä tavoilla, jotka heikentävät yhteiskunnan mahdollisuuksia turvata jäsenilleen mahdollisuudet hyvään, kukoistavaan elämään – seikka, jota lisäkseni muun muassa Nussbaum pitää yhteis-

505 Ks. esimerkiksi Hillyard, P., Pantazis, C., Tombs, S. & Gordon, D. (eds). (2004). *Beyond Criminology: Taking harm seriously*. London: Pluto Press;
Pemberton, S. (2015). Harmful Societies: Understanding Social Harm. Policy Press.

506 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

507 Giddens, A. (1984). The constitution of society: Outline of the theory of structuration: Elements of the theory of structuration. Polity Press.

kunnan tehtävänä[508]. Haitat voivat rapauttaa esimerkiksi oikeusvaltion perustuksia, demokratiaa tai yhteiskunnan toimintamahdollisuuksia. Nämä ovat välttämättömiä, jotta yhteiskunta voisi turvata jäsenilleen puitteet terveelliseen ja turvalliseen elämään. Olennaista on kuitenkin ymmärtää, että myös yhteiskuntaan kohdistuvista haitoista kärsivät lopulta yhteiskunnan jäsenet, joiden mahdollisuudet inhimilliseen kukoistukseen ja hyvään elämään heikkenevät yhteiskunnallisten muutosten seurauksena. Vaikka algoritmisen teknologian haitallinen vaikutus voi kohdistua yhteiskuntaan, sen seuraukset, konkreettiset haitat, palautuvat aina takaisin yksilöihin. Yhteiskunnallisten haittojen tutkimusalan laajentaminen siten, että tällaiset haitat voidaan huomioida jo ensivaiheessa, yhteiskuntaan kohdistuessaan, on merkityksellistä, sillä sen avulla on mahdollista syventää ymmärrystä siitä, millaiset kehityskulut johtavat siihen, että yhteiskunnasta tulee entistä haitallisempi[509] – ja sen seurauksena voi olla mahdollista tunnistaa, minkälaisten muutosten kautta yhteiskuntien haitallisuutta voidaan vähentää.

Välittömien algoritmisten haittojen sijaan yhteiskuntaan vaikuttaisi kohdistuvan lähinnä algoritmisia seurannaishaittoja. Tallaisista haitoista on vasta viime aikoina alettu puhua. Haittojen ilmenemiseen vaikuttavat niin mikro-, meso- kuin makrotason ilmiöt, joista yhä suurempaa osaa algoritmiset teknologiat määrittävät tavalla tai toisella. Haitalliset muutokset voivat olla yllättäviä, epäselviä, vaihtelevia tai vaikeasti havaittavissa. Niihin kuitenkin johtavat kehityskulut, joita on jokseenkin laajasti tunnistettu ja tutkittu: algoritmisten teknologioiden lisääntyvät vaikutukset eri tahojen toiminnan mahdollisuuksiin[510] ja käsityksiin tarpeellisesta ja toivottavasta toiminnasta[511], tekoälyn ja automaation yleistyminen niin työelämässä[512] kuin yhteiskunnallisissa instituutioissa[513], sekä algoritmisten teknologioiden ruokkimat muutokset sosiaalisessa kanssakäymisessä, yhteiskunnallisissa normeissa, käsityksissä merkityksellisestä tiedosta ja osaamisesta, sivistyksessä, kulttuurissa, tavoissa ja ajattelussa[514]. Näiden prosessien seurauksena yhteiskunnan instituutioiden toiminta-

508 Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

509 Yhteiskuntamuotojen haitallisuuden tutkimus kuuluu yhteiskunnallisten haittojen tutkimuksen alueelle. Ks. Pemberton, S. (2015). *Harmful Societies: Understanding Social Harm*. Policy Press.

510 Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.

511 Rosa, H. (2013). *Social Acceleration*. Columbia University Press. 116.

512 Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: how technology changes labor demand. *Journal of Economic Perspectives*, *33*(2), 3–30.

513 Koulu, A. R. (2018). Digitalisaatio ja algoritmit – oikeustiede hukassa? *Lakimies*, *116*(7–8), 840–867.

514 Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford.

edellytykset ja toiminnan tavat muuttuvat, ja osa muutoksista on väistämättä haitallisia.

Lienee todennäköistä, että vasta hyvin pieni osa algoritmisista seurannaishaitoista on tunnistettu[515], puhutaan sitten yksilöihin, ryhmiin tai yhteiskuntaan kohdistuvista seurannaishaitoista. Lisäksi on mahdollista, että joidenkin tunnistettujen, haitallisten ilmiöiden synnyssä algoritmisten teknologioiden roolia ei ole vielä päästy täysimääräisesti arvioimaan. Kuten myös ensimmäisessä osatutkimuksessa[516] argumentoidaan, algoritmit toimivat usein näkymättömissä tai niiden toimintamekaniikasta ei ole tietoa saatavilla, mikä vaikeuttaa niiden haittapotentiaalin ja mahdollisesti realisoituneiden haitallisten vaikutusten arvioimista.

6.2.2 Tekoälylainsäädäntö ja algoritmisten haittojen hallinta

Tähän asti lähinnä suuryrityksillä on ollut laajat mahdollisuudet kehittää algoritmisia teknologioita, joita on voitu hyödyntää jokseenkin vapaasti erilaisissa konteksteissa. Kun järjestelmien toiminnan reunaehtoja tai tuotantoa ei ole säännelty demokraattisten prosessien kautta, yritykset ovat voineet laatia omat eettiset ohjeistuksensa[517], mikä on väistämättä heijastunut järjestelmiin. Niiden kehitystä on ohjannut moderneissa yhteiskunnissa välttämätön tarve kiihtyvällä tahdilla lisätä voittoja, tehokkuutta ja vauhtia[518]. On todennäköistä, että tuottajien mielenkiinto on kohdistunut pääosin sellaisiin riskeihin, jotka altistavat järjestelmien käyttäjät välittömille vahingoille ja järjestelmien tuottajat vahingoista seuraaville imagotappioille tai oikeustaisteluille. Käytännössä merkityksellisiä seikkoja ovat olleet erityisesti tuoteturvallisuus ja tuotteiden houkuttelevuuteen liittyvät seikat. Järjestelmien potentiaali laajaalaisiin, teknologiasta kauempana ilmeneviin seurannaishaittoihin lienee jäänyt vaille teknologioiden tuottajien suurempaa kiinnostusta ainakin osittain siksi, että tällaiset haitat ovat langenneet hyvin harvoin tuottajien vastuulle.

Kansallisesti nimenomaista tekoälysääntelyä on vasta vähän: tekoälyn käyttöä on säännelty toistaiseksi vain hallinnollisen päätöksenteon kontekstissa, jossa automaattista päätöksentekoa ohjaa hallintolain uusi 8 b luku. Sen mukaan hallinnollisessa päätöksenteossa sallitaan ainoastaan sääntöpohjainen automaatio, johon liittyvät riskit ja niihin kiinnittyvä välittömien haittojen potentiaali ovat huomattavasti

515 Ks. myös Yeung, K. (2018). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *MSI-AUT (2018)*, *5.* 62.

516 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

517 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.102–103.

518 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

pienemmät kuin koneoppivien järjestelmien. Kansallisen sääntelyn tiukka rajaus estänee siis tehokkaasti osan tekoälyjärjestelmien potentiaalisesti aiheuttamista haitoista: sääntöpohjaisia monimutkaisempien järjestelmien virheistä kumpuavat välittömät haitat julkisen sektorin päätöksenteossa.

Teknologisten rajausten lisäksi hallinnollisesta automaatiosta mahdollisesti seuraavia haitallisia vaikutuksia todennäköisesti vähentävät lainsäädännön vaatimukset viranomaisprosesseista, joita automaation tarkoituksenmukaisen toiminnan varmistamiseksi tulee olla, sekä automaattisten päätösten tapauksessa vaadittu kattava mahdollisuus oikaisuvaatimusten esittämiseen. Seurannaishaittoja sääntelyssä ei kuitenkaan tunnisteta. Kun sääntely koskee ainoastaan lopullisen hallintopäätöksen tekemistä, se ei tavoita eikä rajoita sellaisten tekoälyjärjestelmien käyttöä, joita hyödynnetään julkisen sektorin päätöksenteon tukena. Työkalut tällaisista järjestelmistä kumpuavien haittojen hallitsemiseksi jäävät vähäiseksi. Sääntelyn rajautuminen ainoastaan julkishallinnolliseen toimintaan luonnollisesti jättää suurimman osan tekoälyjärjestelmistä ja niihin kiinnittyvistä haitoista sääntelyn tarjoamien hallintamahdollisuuksien ulkopuolelle.

EU:n tekoälylainsäädäntö tuo tekoälyjärjestelmät kattavasti sääntelyn piiriin. Tekoälysäädös on luonteeltaan täysharmonisoiva eli se koskee Suomea sellaisenaan. Sääntelyssä tunnistetaan monia tekoälyyn liittyviä, merkityksellisiä riskejä ja määritellään tuottajille ja käyttöönottajille erilaisia, riskitason perusteella määräytyviä velvoitteita. Sääntely ei kuitenkaan koske sääntöpohjaista automaatiota, jollainen on Suomen hallinnollisessa päätöksenteossa sallittu, sillä sääntöpohjaiset järjestelmät eivät ole tekoälyasetuksen määritelmän mukaista tekoälyä. On vielä epäselvää, missä määrin asetuksen voimaantulon jälkeen on kansallisesti mahdollista asettaa tekoälysovellusten käytölle asetusta tiukempia rajoituksia. Käytännössä EU:n tekoälyasetus voi estää hallinnon automaatiomahdollisuuksien rajaamisen ainoastaan sääntöpohjaisilla järjestelmillä toteutettaviksi. Tämä antaisi nykyistä enemmän tilaa riskialttiille, haitallisille käytänteille.

Riskitasoon kiinnittyvän sääntelykokonaisuuden lähtökohtana on ajatus siitä, että tekoälyjärjestelmien riskitaso vaihtelee, ja vaadittujen prosessien laajuuden tulee vaihdella riskitason mukaisesti. Sääntelyn lähtökohta on ymmärrettävä, mutta sen toteutus vaikuttaa kömpelöltä. Asetuksessa määritellään riskikategoriat, joiden perusteella tuottajan ja käyttöönottajan velvoitteet määräytyvät. Suurin osa tekoälysovelluksista lukeutuu rajatun tai pienen riskin sovelluksiin, joita säännellään kevyesti. Kiellettyjen sovellusten lista jää melko suppeaksi, kun innovaatioiden ja markkinoiden rajoittamista vältetään kenties liikaakin. Suuren riskin sovellukset määritellään asetuksen liitteissä I (harmonisoidun EU-lainsäädännön alalla käytettävät tekoälyjärjestelmät tietyin edellytyksin) ja III (järjestelmän käyttötarkoituksen perusteella suuririskisiksi arvioitavat). Jotta sääntely pysyisi relevanttina myös teknologioiden kehittyessä, riskikategorioiden tulisi mukautua teknologiseen kehitykseen. Ennalta

määritelty lista käyttötarkoituksista, jotka arvioidaan suuririskisiksi, voi heikentää mahdollisuuksia reagoida joustavasti tekoälyteknologioiden kehitykseen ja todennäköisesti ennakoimattomalla tavalla lisääntyviin käyttötapoihin. Tuottajien itsearviointiin vahvasti nojaava sääntely voi olla riski itsessään etenkin, kun suuren riskin sovelluksiin liittyvät velvoitteet ovat huomattavasti raskaammat kuin sitä vähäisemmän riskin sovellusten.

Haittojen tunnistaminen ja niihin puuttuminen jää asetuksessa riskienhallinnan varjoon. Etenkin yhteiskuntaan kohdistuvia haittoja on otettu huomioon vain rajallisesti[519], mihin lienee vaikuttanut teknologiayritysten aktiivinen lobbaus[520]. Myös teknologioista etääntyvät seurannaishaitat jäävät sääntelyssä pitkälti tunnistamatta, ja lukuisat rajoitukset pehmentävät suuririskisiin sovelluksiin liitettyjä, ensisilmäyksellä kunnianhimoiselta vaikuttavia velvoitteita[521]. Parlamentti kuitenkin onnistui tuomaan sääntelyyn vakuuttavuutta muun muassa lisäämällä riskienhallintaa painottavien elementtien rinnalle velvoitteen arvioida tekoälyjärjestelmän vaikutuksia: suuren riskin tekoälyjärjestelmien käyttöönottajilla, mikäli nämä toimivat julkishallinnon alalla, on velvollisuus arvioida ihmisoikeusvaikutuksia sovelluksen käyttöympäristössä ja tehdä suunnitelma haitallisiin vaikutuksiin varautumiseksi (art. 27)[522]. Tällöin suuren riskin tekoälysovelluksen käyttöönottajan on arvioitava, minkälaisia kohtuudella ennakoitavissa olevia vaikutuksia järjestelmän käyttöönotolla olisi ihmisoikeuksien kannalta erityisesti haavoittuvassa asemassa olevien ryhmien asemaan.

Vaikka artikla voi auttaa tunnistamaan ja tiedostamaan järjestelmien haittapotentiaalin komission alkuperäistä ehdotusta paremmin, artiklan merkitystä haittojen hallinnassa rajoittaa useampi seikka. Ensinnäkin arviointivelvoitteen ulkopuolelle jätetään esimerkiksi ympäristöön ja yhteiskuntaan kohdistuvat vaikutukset. Toisekseen velvoite on jokseenkin epämääräisesti rajattu kohtuudella ennakoitavissa oleviin vaikutuksiin, ja se koskettaa ainoastaan julkishallinnollisia toimijoita. Kolmanneksi velvoite on riippuvainen tekoälyjärjestelmän arvioidusta riskitasosta. Ihmisoikeusvaikutukset on arvioitava vain, mikäli järjestelmä on määritelty suuririskiseksi.

519 Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).

520 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books. 103;
Schyns, C. (2023). *The lobbying ghost in the machine.* Corporate Europe Observatory.

521 Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 98–99.

522 Parlamentti esitti velvoitetta kaikille suuririskisten tekoälyjärjestelmien käyttöönottajille, mutta lopullisessa versiossa se rajattiin vain julkishallinnollisiin toimijoihin. Ks. Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Muutos 413.

Teknologisen kehityksen kiihtyminen voi jättää uusia, riskialttiita sovelluksia sääntelyn ulkopuolelle. Lisäksi, kun riskitason arvioinnissa nojataan lähtökohtaisesti tuottajan itsearviointiin, on olemassa riski siihen, että joidenkin sovellusten riskitaso arvioidaan opportunistisesti tai tahattomasti mutta virheellisesti todellista pienemmäksi. Tällöin myöskään käyttöönottajalle ei muodostuisi velvoitetta ihmisoikeusvaikutusten arviointiin. Vaikka jälkikäteen riskitason virheellinen arviointi huomattaisiin, haittojen estämisen tai rajoittamisen kannalta se voi olla liian myöhään.

Artikla olisi huomattavasti vaikuttavampi, mikäli se koskisi kaikkia suuren riskin tekoälyjärjestelmien käyttöönottajia, kuten parlamentti esitti. Vielä tehokkaammin haittoja voitaisiin tunnistaa ennakolta, mikäli ihmisoikeusvaikutusten arviointivelvoite laajennettaisiin koskemaan kaikkia tekoälysovelluksia riskitasosta riippumatta. Lisäksi riskitason itsearvioinnissa mahdollisesti tehtyjen virheiden kertautuminen tekoälysovellusten elinkaarten aikana voitaisiin estää. Velvoite ei todennäköisesti muodostuisi suhteettoman raskaaksi: mikäli tekoälysovellusta käytettäisiin rajatussa ympäristössä ja sen vaikutusalue jäisi kapeaksi, vaikutusarvioinnissa huomioitavia seikkoja olisi todennäköisesti kohtuullinen määrä. Jos taas järjestelmän käyttö- ja/tai vaikutusalue olisi laaja, haittapotentiaali ja riskit oletettavasti kasvaisivat, jolloin myös lisääntyvä työmäärä vaikutusten arvioimiseksi olisi perusteltavissa.

Ihmisoikeusvaikutusten arviointivelvoitteen merkittävyyttä toisaalta rajoittaa sen tiukka kiinnittyminen ihmisoikeuskehykseen. Parlamentin kannassa, artiklassa 28 b parlamentti esitti niin kutsuttujen perustamallien – jotka vastaavat melko tarkasti lopullisen säädöksen yleiskäyttöisiä tekoälymalleja – tuottajille velvoitetta arvioida ja rajoittaa haitallisia vaikutuksia paitsi ihmisoikeuksille myös terveydelle, turvallisuudelle, ympäristölle, demokratialle ja oikeusvaltioperiaatteelle[523]. Lopullisessa sääntelyssä systeemisesti riskialttiiden tekoälymallien tuottajille vastaava velvoite asetettiinkin, mutta systeemisesti riskialttiiden mallien raja vedettiin niin korkealle, että sen käytännön merkitys jäänee vähäiseksi. Ihmisoikeusvaikutuksista laajemmalle levittäytyvä arviointivelvoite kuitenkin edesauttaisi etenkin yhteiskuntaan kohdistuvien haittojen tunnistamista ja rajoittamista, mihin tekoälyasetus ei tällaisenaan juuri anna työkaluja. Olisikin ollut perusteltua, että artiklan 27 ihmisoikeusvaikutusten arviointivelvoite olisi asetettu laajempana siten, että myös ihmisoikeuskehyksen ulkopuolelle jäävät riskit olisi tuotu sen piiriin, ja velvoite olisi koskenut nyt säädettyä laajempaa tekoälyjärjestelmien käyttöönottajien piiriä. Tällöin mahdollisesti perustavanlaatuisella tavalla yhteiskuntaa muuttavien teknologioiden vaikutukset olisi saatettu kattavammin sääntelyn piiriin. Vaikutus olisi tietenkin vielä suurempi, mikäli vaikutusten arviointivelvoite koskisi myös muita kuin suuren riskin tekoälysovelluksia.

523 Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Muutos 399.

Laajennettu arviointivelvoite olisikin siirtänyt sääntelyn painotusta ennakollisesta, jokseenkin abstraktista riskien hallinnasta kohden konkreettisempaa haittojen hallintaa ja innovaatioiden sääntelyä[524]. Ratkaisu olisi mahdollistanut myös yhteiskunnan kiihtyvän muutoksen huomioimisen: jos järjestelmän vaikutuksia arvioitaisiin suhteessa sen käyttöönottoympäristöön ennen kuin järjestelmä otettaisiin käyttöön, parhaimmillaan myös ajankohtaiset ilmiöt ja niistä seuraavat ajankuvassa merkitykselliset yhteisvaikutukset voisivat nousta esille. Lopullinen sääntely ei kuitenkaan juuri herätä toiveita tällaisen toteutumisesta.

Ihmisoikeusvaikutusten arviointivelvoite lähenee myös sertifiointiprosessia, joskin sertifioinnissa järjestelmän tuottaja vastaisi sertifioinnin toteuttamisesta. Viranomaisvetoista sertifiointia onkin joissakin yhteyksissä ehdotettu tekoälyjärjestelmien turvallisuuden ja asianmukaisuuden varmistamiseksi[525]. Sertifioinnin sijaan EU-sääntelyssä erilaisilla standardeilla tullee olemaan suuri rooli, ja niiden odotetaan täydentävän sääntelyä, joka vielä toistaiseksi on paikoin hyvinkin tulkinnanvarainen. Standardien noudattaminen on kuitenkin lähtökohtaisesti vapaaehtoista ja ainoastaan yksi keino osoittaa tekoälyjärjestelmän säädöksenmukaisuus (art. 40).

Virallinen, viranomaisvetoiseen vaikutusarviointiin perustuva sertifiointijärjestelmä suurimman riskin tekoälysovelluksille voisi olla niiden laajan ja osittain tuntemattoman haittapotentiaalin takia perusteltavissa. Sellaisen toteuttamisen tueksi tarvittaisiin kuitenkin huomattavasti tässä esitettyä laajempaa tutkimusta, jotta asianmukaiset vaatimukset tekoälyjärjestelmille pystyttäisiin määrittelemään. Sertifiointijärjestelmän luominen ja ylläpitäminen vaatisivat myös mittavia panostuksia niin valtioilta kuin EU:lta, ja tekoälytuotteiden kehitystyö ja markkinoille saattaminen voisivat sertifiointiin liittyvien byrokraattisten prosessien takia hidastua merkittävästi – mitä EU:n tekoälyasetuksessa nimenomaisesti pyritään välttämään, jotta vaikutukset innovaatioihin ja yritysten toimintaan jäisivät mahdollisimman vähäisiksi.

6.2.3 Kiihtyvyyden hillitseminen

Modernissa yhteiskunnassa talouskasvu on välttämätöntä, sillä se luo puitteet muutoksille, joita väistämättä tarvitaan yhteiskunnan vakauttamiseksi[526]. Mikäli tämä lähtökohta hyväksytään, teknologisen kehityksen hidastuminen voimakkaasti innovaatioita rajoittavan lainsäädännön seurauksena johtaisi todennäköisesti laaja-alaisiin, epäsuoriin haitallisiin seurauksiin. Tähän premissiin EU:n tekoälysäädöksessä näytetään nojautuvan. Sen sijaan sääntelyssä ei huomioida sitä, että teknologinen

524 Hasselbalch, J. A. (2018). Innovation assessment: governing through periods of disruptive technological change. *Journal of European Public Policy*, *25*(12), 1855–1873.

525 Ks. esimerkiksi Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.

526 Rosa, H. (2013). *Social Acceleration*. Columbia University Press.

kehitys on jo kiihtynyt siihen pisteeseen, että monilla yhteiskunnan osa-alueilla ei ole enää mahdollisuuksia kiristää tahtia vastaavalla tavalla. Tämä johtaa väistämättä lisääntyvään eritahdistumiseen[527], mikä altistaa laajasti erilaisille haitoille, mikäli muilla yhteiskunnan sektoreilla ei löydetä tapoja synkronoitua teknologisen kehityksen tahtiin.

Hyvä taloustilanne tarjoaa teknologioiden kehitykselle edellytykset. Teknologioiden kehittäminen taas onnistuessaan voimistaa talouskasvua ja ruokkii yhteiskunnan vauhdin kiihtymistä. Talouskasvun mahdollistama innovaatiovauhti vaikuttaisi väistämättä kiihdyttävän ohi mahdollisuuksista tehostaa haittojen hallinnan, vähentämisen ja selvittämisen keinoja. Teknologisen kehityksen kiihdytysmahdollisuuksille ei ole vielä nähtävillä päätepistettä, mutta monien, haittojen estämiseksi ja selvittämiseksi luotujen prosessien kiihdyttäminen vaikuttaa jo olevan mahdotonta. Esimerkiksi monet yhteiskunnan suojamekanismit, kuten yksilön oikeusturvakeinot, toimivat yksittäistapausten tarkan arvioimisen kautta. Monimutkaistuvassa yhteiskunnassa myös selvitettäviksi tulevat tapaukset ovat aiempaa monimutkaisempia, jolloin niiden selvittäminen voi vaatia enemmän työtä ja siis myös aikaa. On siis mahdollista, että prosessit hidastuvat, vaikka tapausmäärät pysyisivät samana. Koska algoritmisia teknologioita leimaa systemaattinen toiminta, niiden käytöstä voi kuitenkin seurata myös virheellisen toiminnan systematisoituminen. Tämä voi lisätä huomattavasti virheitä, minkä seurauksena haitat voivat lisääntyä nopeasti ja tapausmäärät kasvaa[528]. Riskinä on, että teknologisen kehityksen kiihtyessä monet prosessit, joilla algoritmisten teknologioiden mahdollisesti aiheuttamia haittoja pyritään korjaamaan, jäävät kehityksestä jälkeen. Teknologinen kehitys karkaa siis kauemmas mahdollisuuksista suojautua siihen kiinnittyviltä haitallisilta ilmiöiltä.

Jotta pystyttäisiin turvaamaan mahdollisuudet yhtäältä hallita ja *ohjata* teknologioiden kehitystä, toisaalta vähentää algoritmisia haittoja, olisi perusteltua pyrkiä *hidastamaan* teknologioiden kehitystä ja käyttöönottoa ja sitä kautta algoritmista transformaatiota, päinvastoin kuin EU:n tekoälyasetuksella pyritään tekemään. Aiemmin esiin nostettu, sääntelyssä omaksuttua laajempi vaikutustenarviointiprosessi olisi voinut edesauttaa vauhdin hallitsemista. Säädöksen lähtökohdaksi valittu pyrkimys olla heikentämättä kilpailua tai rajoittamatta markkinoita kuitenkin priorisoi teknologiateollisuuden nopeita tulovirtoja ylitse haittojen tehokkaan hallitsemisen.

527 Ibid.

528 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

6.3 Johtopäätelmät

Algoritmiset teknologiat aiheuttavat moninaisia haittoja. Ongelmakenttä, johon algoritmiset haitat kiinnittyvät, on kuitenkin huomattavan monimutkainen ja vaikeasti hallittavissa, sillä haitat syntyvät teknologian, ihmisten ja yhteiskunnan sosioekonoteknisten prosessien vuorovaikutussuhteissa. Ongelmakenttää määrittävät siis niin algoritmiset teknologiat kuin teknologioiden käytölle puitteet antavat yhteiskunnalliset reunaehdot. Nämä vaikuttavat siihen, minkälaisia teknologioita kehitetään, minkälaisia haittoja ne voivat tuottaa, minkälaisilla mekanismeilla haitat syntyvät ja keihin haitat kohdentuvat. Haittoihin puuttumiseksi on siis pyrittävä vaikuttamaan paitsi teknologioihin myös niihin puitteisiin, jotka teknologioiden tuotantoa ja käyttöä määrittävät. Ensin on kuitenkin tunnistettava, minkälaisia haittoja algoritmisiin teknologioihin kiinnittyy.

Algoritmisista haitoista parhaiten tunnettuja ovat haitat, jotka kiinnittyvät suoraviivaisesti teknologian teknisiin ominaisuuksiin tai toiminnan logiikkaan. On laajasti tunnistettu, että huonosti tai epätarkoituksenmukaisesti toteutettu järjestelmä tuottaa väistämättä huonoja tuloksia. Riski haittojen syntymiselle kasvaa, kun algoritmiset teknologiat samalla hävittävät tekijöitä, jotka ovat aiemmin suojanneet haitoilta: esimerkiksi hajautettu ihmisvetoinen toiminta muuntuu keskitetyksi, systemaattiseksi koneelliseksi päättelyksi, jolloin päätösperusteiden ja -käytänteiden jatkuva uudelleenarviointi jää puuttumaan[529].

Erityisesti koneoppivin menetelmin toteutettujen automaattisen päätöksentekojärjestelmien riskitekijöitä tunnetaan jo paljon, ja vaikuttaa siltä, että tekoälyn sääntelypyrkimyksissä niihin onnistutaan puuttumaan jokseenkin kattavasti. Muiden haittojen osalta tilanne on heikompi. Etenkin algoritmisista teknologioista etääntyvät seurannaishaitat jäävät vielä laajasti tunnistamatta niin tutkimuksessa kuin tekoälyn sääntelypyrkimyksissäkin. Ymmärrys siitä, miten algoritmisaatio ja tekoälyn laajamittainen käyttö läpi yhteiskunnan muuttaa yhteiskuntia ja niiden jäseniä, on vielä vaillinaista.

On kuitenkin selvää, että tekoälyteknologioihin tiivistyy valtava muutosvoima. Teknologioiden käytön vaikutuksista johtuvia epäsuoria vaikutuksia on alettu tutkia enenevissä määrin vasta viime vuosina. Tutkimus kiinnittyy suurelta osin laajempaan yhteiskunnallisten transformaatioiden tutkimukseen; sen sijaan esimerkiksi yhteiskunnallisten haittojen tutkimusperinteeseen liittyviä tutkimuksia aiheesta ei vielä juurikaan ole. On kuitenkin nähtävissä, että algoritmisten teknologioiden käyttämi-

529 Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

nen muuttaa yksilöitä, instituutioita ja yhteiskuntaa, ja näistä muutoksista voi seurata haittoja, jotka yltävät yksilötasolta yhteiskuntaan.

Algoritmiset teknologiat vaikuttavat monella tasolla esimerkiksi tietoon, ja näistä vaikutuksista seuraa moninaisia haittoja, jotka kohdistuvat niin yksilöihin, ryhmiin kuin yhteiskuntaankin. Yhtäältä algoritmiset järjestelmät vaativat toimiakseen dataa, ja datasta on tullut valuuttaa, joka muuttaa yksilöt ominaisuuksineen myytäviksi datapisteiksi[530]. Data on siis osaltaan korvannut *tiedon*, ja sen perusteella voidaan tehdä esimerkiksi ennusteita, optimointia ja päätöksiä. Yksilöihin ja ryhmiin kohdistuvat välittömät haitat kytkeytyvät voimakkaimmin tälle tasolle: esimerkiksi opetusdatasta johtuviin vinoumiin ja niistä seuraaviin virheellisiin päätöksiin[531]. Toisaalta data myös määrittää koko ajan voimakkaammin sitä, minkälaiseen tietoon ihminen pääsee käsiksi. Personoitu uutissyöte, hakukoneiden painotukset ja sosiaalisen median suositukset perustuvat käyttäjästä kerättyyn dataan ja algoritmien painotuksiin, ja aggressiivisesti sisältöä rajaavat algoritmit voivat johtaa siihen, että yksilön tietämys ja ymmärrys kaventuu tai vääristyy[532]. Seurauksena voi syntyä monenlaisia haittoja. Yksilöön kohdistuvat autonomiahaitat seuraavat luonnollisesti tiedon rajautumisesta ja kuplautumisesta[533]. Vääristyneet käsitykset voivat johtaa myös polarisaation ja ääriajattelun lisääntymiseen yhteiskunnassa[534], mistä seuraavat haitat voivat kohdistua kokonaisiin ryhmiin. Vaikutukset yltävät tätäkin pidemmälle, sillä esimerkiksi demokratian toiminnan kannalta on välttämätöntä, että riittävän suurella osalla yhteiskunnan jäsenistä on riittävän laaja-alainen ja todellisuutta vastaava käsitys maailmasta. Mikäli näin ei ole, demokratia menettää paitsi oikeutuksen myös tosiasialliset toiminnan mahdollisuudet[535]. Suurin haitta kohdistuu yhteiskuntaan, sillä demokraattisten prosessien murentuessa on todennäköistä, että yhteiskunnan mahdollisuudet hillitä eriarvoisuuden lisääntymistä ja epäoikeudenmukaisuuden voimistumista häiriintyvät. Näiden muutosten aikaansaamat haitat heijastuvat takaisin yksilöihin ja ryhmiin.

Kuten yllä kuvattu esimerkki osoittaa, algoritmiset teknologiat läpäisevät yhteiskunnan perustavanlaatuisella tavalla. Tämän takia niiden vaikutusten arvioiminen

530 Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

531 O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

532 Coeckelbergh, M. (2022). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 1–10.

533 Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298–320.

534 Mondon, A., & Winter, A. (2020). Reactionary democracy: How racism and the populist far right became mainstream. Verso Books.

535 Coeckelbergh, M. (2022). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 1–10.

muista yhteiskunnallisista prosesseista täydellisesti erillään on jokseenkin mahdotonta. Esimerkiksi dataa painotetaan tiedon sijasta todennäköisesti suurelta osin siksi, että organisaatioita ajaa jatkuva paine tehostaa päätöksentekoa ja muita toimenpiteitä. Rekistereihin kerätty henkilötieto nopeuttaa tarpeellisen informaation hakemista ja siten edesauttaa tavoitteen saavuttamista, mutta samalla myös pelkistää todellisuutta tavalla, joka häivyttää inhimillisen elämän monivivahteisuuden.

Myös polarisaation lisääntymiseen vaikuttavat niin algoritmisten teknologioiden edesauttama kuplautuminen ja tiedon rajautuminen kuin myös poliittiset voimat, jotka hyötyvät yhteiskunnan epävakaudesta[536]. Haittojen syntymisessä tekijöitä on siis useita, ja näin ollen algoritmisiin teknologioihin kiinnityvien mutta niistä etääntyvien haittojen tunnistaminen voi olla vaikeaa. Haitalliset vaikutukset voivat ilmetä ajallisesti ja maantieteellisesti huomattavan kaukana itse teknologiasta. Koska haitat voivat ilmetä huomattavalla viiveellä, myös ymmärrys tällaisista haitoista ja niiden perusteista lisääntyy hitaasti. Vielä toistaiseksi vähäinen ymmärrys algoritmisten teknologioiden potentiaalista aikaansaada epäsuoria ja teknologioista etääntyviä haittoja lienee syynä siihen, että EU:n tekoälysääntelyssä tällaiset haitat pitkälti sivuutetaan ja keskitytään niiden sijasta sääntelemään tunnistettuja, välittömiin haittoihin linkittyviä riskejä.

Sääntelyssä valittu riskiperustainen malli ohjaa optimistisesti ajattelemaan, että tekoälyteknologioihin kiinnittyvät riskit pystyttäisiin ennakolta tunnistamaan riittävällä tasolla, jolloin myöskään kohtuutonta määrää haittoja ei pääsisi muodostumaan. Tämä on epätodennäköistä. Tekoälyjärjestelmien laaja-alainen arviointivelvoite, jollaista lopullisessa sääntelyssä ei vaadita, voisi paikata sääntelyn puutteita. Se nostaisi inhimillisen kukoistuksen turvaamisen talouskasvun rinnalle sääntelyä määrittäväksi tekijäksi. Tähän sääntelyssä ei kuitenkaan pyritä. Jos – kuten tällä hetkellä vaikuttaa – sääntelyssä ei tunnisteta algoritmisten järjestelmien haittapotentiaalia riittävän laajasti eikä puututa siihen tehokkaasti, lienee varmaa, että vähäisin rajoituksin kiihtyvän teknologisen kehityksen vanavedessä myös algoritmiset haitat lisääntyvät ja voimistuvat, uusia haittoja syntyy kiihtyvällä vauhdilla, ja haitat kumuloituvat ja kertautuvat yhä uusilla tavoilla sosioekonoteknisessä ympäristössä.

Säädöksessä omaksuttu ratkaisu myös vaikuttaisi vahvistavan teknologia-alan suuryritysten valta-asemaa, ja sen seurauksena valtion toiminnan mahdollisuudet etenkin tekoälyteknologioiden käytön hillitsemiseksi kaventuvat. Sääntelyn pyrkimys olla rajoittamatta markkinoita ja kilpailua vaikuttaisi johtaneen siihen, että suurin osa tekoälyteknologioista jätetään ainoastaan kevyesti säännellyksi. Kiihtyvä teknologinen kehitys on jo heikentänyt demokraattisten prosessien mahdollisuuksia hallita kehitystä, ja sääntelyn seurauksena valtioiden mahdollisuuksia puuttua teko-

536 Mondon, A., & Winter, A. (2020). Reactionary democracy: How racism and the populist far right became mainstream. Verso Books.

älyinnovaatioihin rajataan riskitasojen mukaisesti. Tämä voi johtaa siihen, että tosiasialliset mahdollisuudet vaikuttaa demokratian keinoin teknologisen kehityksen suuntaan heikentyvät. Tekoälyinnovaatioiden haittojen ennaltaehkäisy ja innovaatioista seuraavien hyötyjen saattaminen yhteiskunnassa laajasti saataville vaikuttaisivat jäävän laajasti teknologiateollisuuden toimijoiden hyvän tahdon varaan.

Kun tekoälysäädös mahdollistaa laajasti automaation hyödyntämisen EU:ssa, on todennäköistä, että valtioiden toiminnot muovautuvat tukemaan automaatiomahdollisuuksia. Trendin voi havaita jo osatutkimuksesta 2, jossa argumentoimme, että myös Suomessa perustuslailliset velvoitteet vaikuttavat jäävän osaltaan automaatiopyrkimysten varjoon. Myös Koivisto ja muut[537] katsovat, että erityisesti julkishallinnon lisääntyvä automaatio ja siitä seuraava teknologisen kielen ja käytänteiden siirtyminen osaksi oikeutta voi johtaa oikeusvaltioperiaatteen murrokseen, jossa yksiselitteisyys ja automatisoitavuus nousevat lainsäädäntöä määrittäviksi tekijöiksi ja muokkaavat lain aiemmin kielelliseen – ja sellaisena varsin monimutkaiseen – prosessiin perustuvaa identiteettiä.

Kiihtyvyyden logiikasta ja algoritmisesta transformaatiosta seuraavat haitat tiedostavan tekoälysääntelyn avulla transformaation hallittu hidastaminen ja siitä seuraavien haittojen hallitseminen olisi mahdollista, mikäli poliittista tahtoa tähän olisi. Nykymuodossaan tekoälyn sääntely hidastanee teknologioiden alati kiihtyvää kehitystä ja algoritmista transformaatiota kuitenkin vain vähän – ehkä hieman ironisesti toistaiseksi eniten teknologista kehitystä tullee hidastamaan sääntelyn paikoittainen tulkinnanvaraisuus ja epäselvyys. Niin kansallisessa kuin EU-tason sääntelyssä haittojen tunnistaminen jää riittämättömäksi, mikä väistämättä hankaloittaa haittojen hallintaa ja heikentää paitsi yksilöiden mahdollisuuksia saada oikeussuojaa myös ylipäätään oikeusvaltion perusteita. Teknologisen kehityksen vauhdin hillitseminen olisi välttämätöntä, jotta pystyisimme suuntaamaan yhteiskunnan muutoksen algoritmisen transformaation keskellä kohti nykyistä kestävämpää, inhimillisen kukoistuksen paremmin mahdollistavaa yhteiskuntajärjestelmää – tai edes ylläpitämään nykyisen tason haittojen hallinnassa. Tätä varten EU:n tekoälysääntelyssä tulisi kuitenkin huomioida nykyistä laaja-alaisemmin erilaiset tekoälyteknologioihin ja algoritmiseen transformaatioon liittyvät välittömät ja seurannaishaitat, luoda tehokkaita keinoja niiden arvioimiseksi ja ennalta estämiseksi, sekä uskaltaa tarvittaessa rajoittaa haitalliseksi arvioitujen tekoälysovellusten tuotantoa, markkinoille saattamista ja käyttöönottoa.

537 Koivisto, I., Koulu, R., & Larsson, S. (2024). User accounts: How technological concepts permeate public law through the EU's AI Act. *Maastricht Journal of European and Comparative Law*, 1023263X241248469.

Lähteet

(Internet-lähteet merkattu *-symbolilla.)

Acemoglu, D. (2021). *Harms of AI* (No. w29247). National Bureau of Economic Research.

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: how technology changes labor demand. *Journal of Economic Perspectives*, *33*(2), 3–30.

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197.

Akansu, A. N. (2017). The flash crash: a review. *Journal of Capital Markets Studies*.

Algan, Y. (2018). Trust and social capital. *For Good Measure: Advancing Research on Well-Being Metrics Beyond GDP; Stiglitz, J., Fitoussi, J., Durand, M., Eds*, 283–320.

Almond, D., Du, X., & Vogel, A. (2022). Reduced trolling on Russian holidays and daily US Presidential election odds. *Plos one, 17*(3), e0264507.

* Alvesalo-Kuusi, A., & Tolsa, T. (2022). *Mihin kriminologin katse kohdistuu – Vastaamo-tapauksen monet kasvot.* (24.11.2022). *Haaste*. Saatavilla <https://rikoksentorjunta.fi/-/haaste-4-22-mihin-kriminologin-katse-kohdistuu>.

Alvesalo, A. (1999). Meeting the expectations of the local community on safety—what about white-collar crime? Konferenssiesitys *27th Annual Conference of the European Group for the Study of Deviance and Social Control*, Liettua, 2.–5.9.1999.

Ananny, M. (2016). Toward an ethics of algorithms: Convening, observation, probability, and timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117.

* Angwin, J., Tobin, A., & Varner, M. (2017). Facebook (still) letting housing advertisers exclude users by race. (21.11.2017). *ProPublica*. Saatavilla <https:// www.propublica.org/article/facebook- advertising- discrimination- housing- race- sex- national- origin>.

Arendt, H. (1973). The origins of totalitarianism. *New York*.

Ayodele, T. O. (2010). Types of machine learning algorithms. Teoksessa *New advances in machine learning*. Zhang, Y. (Ed.). 19–48.

Bahl, U., Topaz, C. M., Obermüller, L., Goldstein, S., & Sneirson, M. (2023). Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. *UCLA L. Rev. Discourse*, *71*, 246.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Bayamlıoğlu, E., & Leenes, R. (2018). The 'rule of law' implications of data-driven decision-making: a techno-regulatory perspective. *Law, Innovation and Technology,* 10(2), 295–313.

Benito, A. (2006). Does job insecurity affect household consumption? *Oxford Economic Papers*, *58*(1), 157–181.

Bennett, W., & Livingston, S. (Eds.). (2020). *The Disinformation Age* (SSRC Anxieties of Democracy). Cambridge: Cambridge University Press.

Bijker, W. E., & Law, J. (Eds.). (1994). *Shaping technology/building society: Studies in sociotechnical change*. MIT press.

Binns, R. (2022). Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*, *16*(1), 197–211.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. Konferenssijulkaisu, *Proceedings of the 2018 Chi conference on human factors in computing systems* (1–14).

Bird, E., Fox-Skelly, J., Jenner, N., Larbey, R., Weitkamp, E., & Winfield, A. (2020). *The ethics of artificial intelligence: Issues and initiatives*. European Parliamentary Research Service.

Bobric, G. D. (2021). The overton window: A tool for information warfare. In *ICCWS 2021 16th International Conference on Cyber Warfare and Security* (p. 20). Academic Conferences Limited.

Bostrom, N. (2014). *Superintelligence.* Oxford University Press.

Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation. *Computational Propaganda Research Project.*

Brayne, S. (2020). *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press, USA.

Brownsword, R. (2015). In the year 2061: from law to technological management. *Law, Innovation and Technology*, *7*(1), 1–51.

Brownsword, R. (2016). Technological management and the rule of law. *Law, Innovation and Technology*, *8*(1), 100–140.

Brownsword, R. (2019). *Law, technology and society: reimagining the regulatory environment*. Routledge.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*, 273–291.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, *20*(1), 15–26.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, communication & society*, *20*(1), 30–44.

Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology*, *47*, 213–237.

Canning, V. (2019). Abject asylum: Degradation and the deliberate infliction of harm against refugees in Britain. *Justice, Power and Resistance*, *3*(1), 37–60.

Canning, V., & Tombs, S. (2021). *From social harm to zemiology: A critical introduction*. Routledge.

Castets-Renard, C., & Besse, P. (2023). Ex ante Accountability of the AI Act: Between Certification and Standardization, in Pursuit of Fundamental Rights in the Country of Compliance. *Pursuit of Fundamental Rights in the Country of Compliance. Artificial Intelligence Law: Between Sectoral Rules and Comprehensive Regime. Comparative Law Perspectives. C. Castets-Renard & J. Eynard (eds),* Bruylant, Forthcoming.

Chan, J. (2022). Online astroturfing: A problem beyond disinformation. *Philosophy & Social Criticism*, 01914537221108467.

Che, T., Liu, X., Li, S., Ge, Y., Zhang, R., Xiong, C., & Bengio, Y. (2021, May). Deep verifier networks: Verification of deep discriminative models with deep generative models. Konferenssijulkaisu, *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 35, No. 8, 7002–7010.

Chiusi, F., Fischer, S., Kayser-Bril, N., & Spielkamp, N. (2020). Automating society report 2020. AlgorithmWatch.

Cinnamon, J. (2017). Social injustice in surveillance capitalism. *Surveillance & Society*, *15*(5), 609–625.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Wash. L. Rev.*, *89*.

Coeckelbergh, M. (2019). Artificial intelligence: some ethical issues and regulatory challenges. *Technology and regulation*, *2019*, 31–34.

Coeckelbergh, M. (2020). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*(4), 2051–2068.

Coeckelbergh, M. (2022). Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 1–10.

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior*, *97,* 202–215.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Nill Sánchez, A., Raji, D., Lisi Rankin, J., Richardson, R., Schultz, J., Myers West, S., & Whittaker, M. (2019). AI now 2019 report. *AI Now Institute*.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation* (pp. 289–294). Routledge.

Czaja, I., & Urbaniec, M. (2019). Digital exclusion in the labour market in European countries: Causes and consequences. *European Journal of Sustainable Development*, *8*(5), 324–336.

Day, J., Iwańska, K., Simon, E. & Willamo, K. (2024). *Packed with loopholes: Why the AI act fails to protect civic space and the rule of law.* Civil Liberties Union for Europe e.V.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, *11*, 1166120.

Dornberger, R., Inglese, T., Korkut, S., & Zhong, V. J. (2018). Digitalization: Yesterday, today and tomorrow. *Business Information Systems and Technology 4.0: New Trends in the Age of Digital Change*, 1–11.

* Doyle, D. (2020). The Overton Window has been flung wide open. (1.5.2020). *Medium.* Saatavilla <https://medium.com/@daithioduill/the-overton-window-has-been-flung-wide-open-fdef96e07840>.

Eg, R., Tønnesen, Ö. D., & Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, *9*, 100253.

Elliott, A. (2019). *The culture of AI: Everyday life and the digital revolution*. Routledge.

Ellis, J. R. (2022). Blurred consent and redistributed privacy: owning LGBTQ identity in surveillance capitalism. In *Diversity in Criminology and Criminal Justice Studies,* 183–196. Emerald Publishing Limited.

Ellul, J. (1964). *The technological society*. New York: Vintage.

Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153.

Esko, T., & Koulu, R. (2022). Rethinking research on social harms in an algorithmic context. *Justice, Power and Resistance*, *5*(3), 307–313.

Eubanks, V. (2017). *Automating Inequality*. St. Martin's Press.

Euroopan ihmisoikeustuomioistuimen tutkimusosasto. (2013). *National security and European case-law.* Euroopan neuvosto/ECHR.

Executive Office of the President. (2014). *Big Data: Seizing Opportunities, Preserving Values.*

Fairclough, N. (1996). *Discourse and Social Change*. Cambridge: Polity Press.

Fiorentino, V., Harrikari, T., Saraniemi, S., & Romakkaniemi, M. (2023). Sosiaalialan työn kiihtyvyys COVID-19-pandemian seurauksena. *Sosiaalityö kriiseissä.* 243.

* Flanagan, M. (2018). The rise of the "Automacene": How robots will define the next epoch in human history. (16.6.2018). *Salon.* Saatavilla < https://www.salon.com/2018/06/16/the-rise-of-the-automacene-how-robots-will-define-the-next-epoch-in-human-history/>.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, *80*(S1), 298–320.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. OUP Oxford.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). CapAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. SSRN 4064091.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14.* 349–379.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., ... & Rahwan, I. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, *116*(14), 6531–6539.

Friedrichs, D. O. (2009). *Trusted criminals: White collar crime in contemporary society*. Cengage Learning.

* Gary, M. (2023). Inside the Heart of ChatGPT's Darkness (11.2.2023). *Marcus on AI.* Saatavilla <https://garymarcus.substack.com/p/inside-the-heart-of-chatgpts-darkness>.

Gehl, R. W., & Bakardjieva, M. (Eds.). (2016). *Socialbots and their friends: Digital media and the automation of sociality*. Taylor & Francis.

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Cambridge: Polity Press.

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, communication & society*, *20*(1), 63–80.

Glied, S., & Lleras-Muney, A. (2008). Technological innovation and inequality in health. *Demography*, *45*, 741–761.

Green, B., & Chen, Y. (2021). Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–33.

Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artificial Intelligence and Law*, *30*(3), 291–323.

* Grossman, G. (2023). Generative AI may only be a foreshock to AI singularity. (11.2.2023). *VentureBeat.* Saatavilla <https://venturebeat.com/ai/generative-ai-may-only-be-a-foreshock-to-ai-singularity/>.

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access.

Hall, S., & Winlow, S. (2018). Ultra-realism. Teoksessa *Routledge Handbook of Critical Criminology* (2. painos). 43–56. Routledge.

Hallevy, G. (2015). *Liability for crimes involving artificial intelligence systems*. New York, NY, USA: Springer International Publishing.

Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100.

Hannan, J. (2018). Trolling ourselves to death? Social media and post-truth politics. *European Journal of Communication, 33*(2), 214–226.

* Harari, Y. (2023). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. (28.4.2023). *The Economist.* Saatavilla <https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation>.

Hasselbalch, J. A. (2018). Innovation assessment: governing through periods of disruptive technological change. *Journal of European Public Policy*, *25*(12), 1855–1873.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2), 209–231.

* Heleskoski, J. (2023). Äärimmäisen nolo esimerkki tekoälyn vaaroista: emämunauksen tehnyt juristi perustelee tekoaan. (12.6.2023). *Mikrobitti.* Saatavilla <https://www.mikrobitti.fi/uutiset/aarimmaisen-nolo-esimerkki-tekoalyn-vaaroista-emamunauksen-tehnyt-juristi-perustelee-tekoaan/53190932-aae0-4a53-af5c-0ec3dd004ae4>.

Helminen, M., & Alvesalo-Kuusi, A. (2017). Advocating the 'Good' Criminal Justice System. The Involvement and Ideas of Civil Society Organisations in Formulating Finnish Criminal Policy. *Retfærd. Nordic Journal of Law and Justice* 40 (2), 3–24.

Hijazi, S., Kumar, R., & Rowen, C. (2015). Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*, *9*(1).

Hildebrandt, M. (2015). *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing. 22–30.

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170355.

Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (2004). *Beyond criminology: Taking harm seriously*. Pluto Press.

Hillyard, P. & Tombs, S. (2004). Beyond criminology? Teoksessa Hillyard, P., Pantazis, C., Tombs, S., & Gordon, D. (eds) *Beyond criminology: Taking harm seriously*: 10–29. Pluto Press.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology*, *11*, 19–29.

* Hinduja, S. (2023). Generative AI as a Vector for Harassment and Harm. (10.5.2023). Cyberbullying Research Center. Saatavilla <https://cyberbullying.org/generative-ai-as-a-vector-for-harassment-and-harm>.

Hine, E., & Floridi, L. (2023). The Blueprint for an AI Bill of Rights: in search of enaction, at risk of inaction. *Minds and Machines*, 1–8.

Hirvonen, H. (2022). Virkavastuu ja päätösautomaatio – vastuun henkilökohtaisuus kriisissä? *Lakimies*, *2022*(3–4), 386–418.

* Horwitz, J., & Seetharaman, D. (2020). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. (26.5.2020) *The Wall Street Journal*. Saatavilla <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed- solutions-11590507499>.

Ilan, J. (2019). Cultural criminology: The time is now. *Critical Criminology*, *27*, 5–20.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The information society*, *16*(3), 169–185.

Jasanoff, S. (2004). The idiom of co-production. Teoksessa *States of knowledge* (pp. 1–12). Routledge.

Jasanoff, S. (2020). Constitutional moments in governing science and technology. Teoksessa *The Ethics of Nanotechnology, Geoengineering, and Clean Energy* (pp. 477–494). Routledge.

Jensen, S. Q. (2011). Othering, identity formation and agency. *Qualitative Studies*, *2*(2), 63–78.

Jones, H., & Hietanen, J. (2023). The r/wallstreetbets 'war machine': Explicating dynamics of consumer resistance and capture. *Marketing Theory, 23*(2), 225–247.

Jones, M. L. (2018). Does technology drive law? The dilemma of technological exceptionalism in cyberlaw. *U. Ill. JL Tech. & Pol'y*, 249.

Kalla, J. L., & Broockman, D. E. (2018). The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments. *American Political Science Review*, 112/1. 148–166.

Karim, F., Oyewande, A. A., Abdalla, L. F., Ehsanullah, R. C., & Khan, S. (2020). Social media use and its connection to mental health: a systematic review. *Cureus*, *12*(6).

Katzenbach, C., & Ulbricht, L. (2019). Algorithmic governance. *Internet Policy Review*, *8*(4), 1–18.

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. *The British Journal of Criminology*, *59*(3), 674–692.

Kim, Y. J., Kim, K., & Lee, S. (2017). The rise of technological unemployment and its implications on the future macroeconomic landscape. *Futures*, *87*, 1–9.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics*, *26*, 89–120.

* Koivisto, I. (2023). Automaattinen päätöksenteko tulee – oletko valmis? (2.3.2023) *Perustuslakiblogi: Suomen valtiosääntöoikeudellisen seuran ajankohtaispalsta*.
Koivisto, I., & Koulu, R. (2020). Miten hyvä hallinto digitalisoidaan? Haaste oikeustieteelliselle tutkimukselle. *Lakimies*, *118*(6), 798–821.
Koivisto, I., Koulu, R., & Larsson, S. (2024). User accounts: How technological concepts permeate public law through the EU's AI Act. *Maastricht Journal of European and Comparative Law,* 1023263X241248469.
Kong, Y. (2022). Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. Konferenssijulkais, *2022 ACM Conference on Fairness, Accountability, and Transparency.*
Koulu, A. R. (2018). Digitalisaatio ja algoritmit – oikeustiede hukassa? *Lakimies*, *116*(7–8), 840–867.
Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital transformation: An overview of the current state of the art of research. *Sage Open*, *11*(3), 21582440211047576.
Kurzweil, R. (2022). Superintelligence and singularity. In *Machine Learning and the City: Applications in Architecture and Urban Design*, 579–601.
Kuusi, O., & Heinonen, S. (2020). Tulevaisuuspolkuja kapeasta tekoälystä vahvaan tekoälyyn. *Tieteessä tapahtuu*, *38*(3).
König, P. D., & Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. *Technology in Society*, *67*, 101688.
Lasslett, K. (2010). Crime or social harm? A dialectical perspective. *Crime, Law and Social Change*, *54*, 1–19.
Latzer, M. (2009). Information and communication technology innovations: radical and disruptive?. *New Media & Society*, *11*(4), 599–619.
Lavorgna, A., & Ugwudike, P. (2022). Managing risks, passing over harms? A commentary on the proposed EU AI Regulation in the context of criminal justice. *Justice, Power and Resistance*, *5*(3), 292–298.
Ledwich, M., & Zaitsev, A. (2019). Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211*.
* Lee, S. (2020). Our Short-Lived Anthropocene and the Coming Algorithmocene. (20.9.2020). Medium. Saatavilla <https://medium.com/swlh/our-short-lived-anthropocene-and-the-coming-algorithmocene-edaf0a07f534>.
Leiser, F., Eckhardt, S., Knaeble, M., Maedche, A., Schwabe, G., & Sunyaev, A. (2023). From ChatGPT to FactGPT: A participatory design study to mitigate the effects of large language model hallucinations on users. *Proceedings of Mensch und Computer 2023* (81–90).
Lepinkäinen, N., & Malik, H. M. (2022). Discourses on AI and regulation of automated decision-making. *Global Perspectives*, *3*(1), 33707.
Lin, T. C. (2016). The new market manipulation. *Emory LJ*, *66*, 1253.
Makhortykh, M., Urman, A., & Ulloa, R. (2021). Detecting race and gender bias in visual representation of AI on web search engines. *Advances in Bias and Fairness in Information Retrieval: Second International Workshop on Algorithmic Bias in Search and Recommendation, BIAS 2021, Lucca, Italy, April 1, 2021, Proceedings* (36–50). Cham: Springer International Publishing.
Malik, H. M., & Lepinkäinen, N. (2022). Between algorithmic and analogue harms: the case of automation in Finnish Immigration Services. *Justice, Power and Resistance*, *5*(3).
Malik, H. M., Lepinkäinen, N., Alvesalo-Kuusi, A., & Viljanen, M. (2022). Social harms in an algorithmic context. *Justice, Power and Resistance*, *5*(3), 193–207.
Malik, H. M., Viljanen, M., Lepinkäinen, N., & Alvesalo-Kuusi, A. (2022). Dynamics of social harms in an algorithmic context. *International Journal for Crime, Justice and Social Democracy*, *11*(1), 182–195.

Manninen, O. (2016). Markkinalikviditeetistä, sen merkityksestä ja kestävyydestä. Suomen pankin ajankohtaisia artikkeleita taloudesta. Saatavilla <https://www.eurojatalous.fi/fi/2016/artikkelit/markkinalikviditeetista--sen-merkityksesta-ja-kestavyydesta/>.

Mann, M. (2020). Technological politics of automated welfare surveillance: Social (and data) justice through critical qualitative inquiry. *Global Perspectives*, *1*(1).

Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., ... & Siemens, G. (2022). Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI? *Computers and Education: Artificial Intelligence*, *3*, 100056.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mazur, O. (2018). Taxing the robots. *Pepp. L. Rev.*, *46*, 277.

McCarthy, D. R. (2013). Technology and 'the international'or: How I learned to stop worrying and love determinism. *Millennium*, *41*(3), 470–490.

McQuillan, Dan. *Resisting AI: an anti-fascist approach to artificial intelligence*. Policy Press, 2022.

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. *Psychological bulletin*, *129*(5), 674.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*/2, 205395171667967.

Mondal, S., Das, S., & Vrana, V. G. (2023). How to bell the cat? A theoretical review of generative artificial intelligence towards digital disruption in all walks of life. *Technologies*, *11*(2)

Mondon, A., & Winter, A. (2020). *Reactionary democracy: How racism and the populist far right became mainstream*. Verso Books.

Moses, L. B. (2007). Recurring dilemmas: The law's race to keep up with technological change. *U. Ill. JL Tech. & Pol'y*, 239.

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.

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach* (Vol. 3). Cambridge University Press.

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ojanen, A., Sahlgren, O., Vaiste, J., Björk, A., Mikkonen, J., Kimppa, K., Laitinen, A. & Oljakka, N. (2022). Algoritminen syrjintä ja yhdenvertaisuuden edistäminen: Arviointikehikko syrjimättömälle tekoälylle. Valtioneuvoston kanslia.

Ossewaarde, M., & Gulenc, E. (2020). National varieties of artificial intelligence discourses: Myth, utopianism, and solutionism in West European policy expectations. *Computer*, *53*(11), 53–61.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Pemberton, S. (2015). *Harmful Societies: Understanding Social Harm*. Policy Press.

Persily, N., & Tucker, J. A. (Eds.). (2020). Social media and democracy: The state of the field, prospects for reform.

* Petkauskas, V. (2023). ChatGPT's answers could be nothing but a hallucination. (6.3.2023). *Cybernews.* Saatavilla <https://cybernews.com/tech/chatgpts-bard-ai-answers-hallucination/>.

Pirjatanniemi, E., Lilja, I., Helminen, M., Vainio, K., Lepola, O., & Alvesalo-Kuusi, A. (2021). Ulkomaalaislain ja sen soveltamiskäytännön muutosten yhteisvaikutukset kansainvälistä suojelua hakeneiden ja saaneiden asemaan. *Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja* 2021:10.

Pöysti, T. H. (2018). Kohti digitaalisen ajan hallinto-oikeutta. *Lakimies*, *2018*(7–8), 868–903.

Raley, R. & Rhee, J. (2023). Critical AI: A Field in Formation. *American Literature*, 95(2). 185–204.
Raymen, T. (2022). *The enigma of social harm: The problem of liberalism*. Taylor & Francis.
* Rochier. (2024). The European Parliament adopts the Artificial Intelligence Act. (29.4.2024). *Rochier Insights.* Saatavilla <https://www.roschier.com/newsroom/the-european-parliament-adopts-the-artificial-intelligence-act>.
Rommetveit, K., & Van Dijk, N. (2022). Privacy engineering and the techno-regulatory imaginary. *Social Studies of Science*, *52*(6), 853–877.
Rosa, H. (2010). *Alienation and Acceleration. Towards a Critical Theory of Late-Modern Temporality.* NSU Press.
Rosa, H. (2013). *Social Acceleration*. Columbia University Press.
Rosa, H. (2015). Escalation: The crisis of dynamic stabilisation and the prospect of resonance. Sociology, capitalism, critique.
Rosa, H. (2017). Available, accessible, attainable: The mindset of growth and the resonance conception of the good life. Teoksessa *The Good life beyond growth* (39–53). Routledge.
Rosa, H. (2018). Airports built on shifting grounds? Social acceleration and the temporal dimension of law. Teoksessa *Temporal Boundaries of Law and Politics* (pp. 72–87). Routledge.
Rosa, H., Dörre, K., & Lessenich, S. (2017). Appropriation, activation and acceleration: The escalatory logics of capitalist modernity and the crises of dynamic stabilization. *Theory, Culture & Society*, *34*(1), 53–73.
Rothe, D., & Kauzlarich, D. (2016). *Crimes of the powerful: An introduction*. Routledge.
Russell, S. (2019). It's not too soon to be wary of AI: We need to act now to protect humanity from future superintelligent machines. *IEEE Spectrum*, *56*(10), 46–51.
Ruuska, T., & Heikkurinen, P. (2023). Kokoava ote teknologiaan Marxiin ja Heideggeriin pohjautuen. *Tutkimus & kritiikki*, *3*(1), 68–87.
Rothstein, B., & Stolle, D. (2008). The state and social capital: An institutional theory of generalized trust. *Comparative politics*, *40*(4), 441–459.
Sahbaz, U. (2019). Artificial intelligence and the risk of new colonialism. *Horizons: Journal of International Relations and Sustainable Development*, (14), 58–71.
Sanders, C. K., & Scanlon, E. (2021). The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Journal of Human Rights and Social Work*, *6*, 130–143.
Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118.
Saurwein, F., & Spencer-Smith, C. (2021). Automated trouble: The role of algorithmic selection in harms on social media platforms. *Media and Communication*, *9*(4), 222–233.
Schyns, C. (2023). *The lobbying ghost in the machine.* Corporate Europe Observatory.
Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278.
Shaefer, H.L., & Grey, S. (2015). Letter to U.S. Department of Labor – Michigan Unemployment Insurance Agency: Unjust Fraud and Multiple-Determinations. Saatavilla <http://www.uiafraudclassaction.com/wp-content/uploads/2017/08/bauserman-u-of-m-memo-to- dol-1.pdf>
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 (TiiS)*, *10*(4), 1–31.
Shoss, M. K. (2017). Job insecurity: An integrative review and agenda for future research. *Journal of management*, *43*(6), 1911–1939.
Sismondo, S. (2017). Post-truth? *Social studies of science, 47*(1), 3–6.
Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). ChatGPT: More than a Weapon of Mass Deception, Ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. *arXiv preprint arXiv:2304.11215*.

Smith, G. J., Bennett Moses, L., & Chan, J. (2017). The challenges of doing criminology in the big data era: Towards a digital and data-driven approach. *The British journal of criminology*, *57*(2), 259–274.

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve legally trustworthy AI: a response to the European Commission's proposal for an artificial intelligence act. *SSRN 3899991*.

Smuha, N. A. (2021). Beyond the individual: governing AI's societal harm. *Internet Policy Review*, *10*(3).

Smuha, N. A. (2021). From a 'race to AI'to a 'race to AI regulation': regulatory competition for artificial intelligence. *Law, Innovation and Technology*, *13*(1), 57–84.

Snider, L. (2014). Interrogating the algorithm: Debt, derivatives and the social reconstruction of stock market trading. *Critical Sociology*, *40*(5), 747–761.

Soliman, F. (2021). States of exception, human rights, and social harm: Towards a border zemiology. *Theoretical Criminology*, *25*(2), 228–248.

Solum, L. B. (1991). Legal personhood for artificial intelligences. *NCL Rev., 70*, 1231.

Spaulding, N. W. (2020). Is Human Judgment Necessary?: Artificial Intelligence, Algorithmic Governance, and the Law. In *The Oxford Handbook of Ethics of AI*. Oxford University Press.

Stark, B., Stegmann, D., Magin, M., & Jürgens, P. (2020). Are algorithms a threat to democracy? The rise of intermediaries: A challenge for public discourse. *AlgorithmWatch.*

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48/2. 204–231.

Stones, R. (2017). *Structuration theory*. Bloomsbury Publishing.

Stratton, G., Powell, A., & Cameron, R. (2017). Crime and justice in digital society: Towards a 'digital criminology'? *International Journal for Crime, Justice and Social Democracy*, *6*(2).

Stuurman, K., & Lachaud, E. (2022). Regulating AI. A label to complete the proposed Act on Artificial Intelligence. *Computer Law & Security Review*, *44*, 105657.

Sumpter, D. (2018). *Outnumbered: From Facebook and Google to Fake News and Filter-bubbles - the algorithms that control our lives*. Bloomsbury Publishing.

Sunstein, C. (2018). # *Republic*. Princeton university press.

Sunstein, C. R. (2019). *Conformity*. New York University Press.

Susser, D. (2022, June). Decision Time: Normative Dimensions of Algorithmic Speed. In *2022 ACM Conference on Fairness, Accountability, and Transparency.*

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American sociological review*, *22*(6), 664–670.

Taulli, T. (2023). Large Language Models: How Generative AI Understands Language. Teoksessa *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business* (93–125). Berkeley, CA: Apress.

Tegmark, M. (2018). *Life 3.0: Being human in the age of artificial intelligence*. Vintage.

Thaler, R. H. (2018). Nudge, not sludge. *Science*, *361*/6401.

Timmermans, S., & Kaufman, R. (2020). Technologies and health inequities. *Annual Review of Sociology*, *46*, 583–602.

Todolí-Signes, A. (2019). Algorithms, artificial intelligence and automated decisions concerning workers and the risks of discrimination: the necessary collective governance of data protection. *Transfer: European Review of Labour and Research*, *25*(4), 465–481.

Tombs, S. ja Hillyard, P. (2004). Towards a political economy of harm: States, corporations and the production of inequality. In Hillyard P, Pantazis C, Tombs S and Gordon D (eds) *Beyond Criminology: Taking harm seriously*: 30–54. London: Pluto Press.

Tombs, S. (2019). Grenfell: the unfolding dimensions of social harm. *Justice, Power and Resistance 3/1*. 61–88.

Tufekci, Z. (2015). Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Journal on Telecommunications & High Tech Law*, *13*/23. 203–216.

Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Springer.
Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.
U.S. Commodity Futures Trading Commission, & U.S. Securities & Exchange Commission. (2010). *Finding regarding the market events of May 6. 2010. Report of the staffs of the CTFC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*.
Vainio, N., Tarkka, V., & Jaatinen, T. (2020). *Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista.* Oikeusministeriön julkaisuja, Selvityksiä ja ohjeita 2020:14.
Van den Bos, K. (2020). Unfairness and radicalization. *Annual review of psychology*, *71*, 563–588.
Van der Wagen, W., & Pieters, W. (2015). From cybercrime to cyborg crime: Botnets as hybrid criminal actor-networks. *British journal of criminology*, *55*(3), 578–595.
Van Dijk, J. (2020). *The digital divide*. John Wiley & Sons.
Van Parijs, P. (2004). Basic income: a simple and powerful idea for the twenty-first century. *Politics & Society*, *32*(1), 7–39.
Van Prooijen, J. W., Spadaro, G., & Wang, H. (2022). Suspicion of institutions: How distrust and conspiracy theories deteriorate social relationships. *Current opinion in psychology*, *43*, 65–69.
Vanto, J., Saarikkomäki, E., Alvesalo-Kuusi, A., Lepinkäinen, N., Pirjatanniemi, E., & Lavapuro, J. (2022). Collectivized Discretion: Seeking Explanations for Decreased Asylum Recognition Rates in Finland After Europe's 2015 "Refugee Crisis". *International Migration Review*, *56*(3), 754–779.
Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112.
Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. Teoksessa *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm,* 175–191. Cham: Springer International Publishing.
Viljanen, M. (2017). Algoritmien haaste: uuteen aineelliseen oikeuteen? *Lakimies 115 (2017): 7–8,* 1070–1087.
Viljanen, M. (2022). Technology matters: how algorithm and artificial intelligent technology features affect harms reduction efforts. *Justice, Power and Resistance*, *5*(3), 314–321.
Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.
Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567.
Wall, D. (Ed.). (2001). *Crime and the Internet*. Routledge.
Whyte, D. (2017). Crime as a social relation of power: Reframing the 'ideal victim' of corporate crimes. Teoksessa *Handbook of Victims and Victimology* (pp. 333–347). Routledge.
Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy*, *25*(6), 865–899.
Willson, M. (2019). Algorithms (and the) everyday. Teoksessa *The Social Power of Algorithms* (137–150). Routledge.
Wodak, R., & Meyer, M. (eds). (2001). *Methods of Critical Discourse Analysis*. London: SAGE Publications, Ltd.
Wodak, R. (2015). Critical Discourse Analysis, Discourse-Historical Approach. Teoksessa *The International Encyclopedia of Language and Social Interaction*, 1–14.
Wood, M. A. (2021). Rethinking how technologies harm. *The British Journal of Criminology*, *61*(3), 627–647.
Yar, M. (2012). Critical criminology, critical theory and social harm. Teoksessa S. Hall, S. Winlow (eds). *New Directions in Critical Theory*. 52–63. Routledge.
Yeung, K., & Lodge, M. (Eds.). (2019). *Algorithmic regulation*. Oxford University Press.

Yeung, K. (2011). Can we employ design-based regulation while avoiding brave new world? *Law, Innovation and Technology*, *3*(1), 1–29.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118–136.

Yeung, K. (2018). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *MSI-AUT (2018)*, *5*.

Yeung, K. (2019). Responsibility and AI - A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. Council of Europe. Saatavilla <https://rm.coe.int/responsability-and-ai-en/168097d9c5>.

Yeung, K. (2019). Why worry about decision-making by machine? Teoksessa *Algorithmic regulation*. Oxford University Press.

* Yle. (2017). Report: Immigration Service circulates model negative asylum decisions for 'assembly line' use. (4.5.2017). Saatavilla <https://yle.fi/news/3-9594858>.

* Yle. (2023). Facebookin moderaattorit vaativat oikeuksiaan ja saivat potkut Keniassa – "Facebook on uusi siirtomaavalta". (10.5.2023). Saatavilla <https://yle.fi/a/74-20030960>.

Young, M. M., Bullock, J. B., & Lecy, J. D. (2019). Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration. *Perspectives on Public Management and Governance*, *2*(4), 301–313.

* Zagni, G., & Canetta, T. (2023). Generative AI marks the beginning of a new era for disinformation. (5.4.2023). *European Digital Media Observatory.* Saatavilla <https://edmo.eu/2023/04/05/generative-ai-marks-the-beginning-of-a-new-era-for-disinformation/>.

Zalnieriute, M., Moses, L. B., & Williams, G. (2019). The rule of law and automation of government decision-making. *The Modern Law Review*, *82*(3), 425–455.

Zarouali, B., Helberger, N., & De Vreese, C. H. (2021). Investigating algorithmic misconceptions in a media context: Source of a new digital divide? *Media and Communication*, *9*(4), 134–144.

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology*, *18*(5), 623–642.

Zeng, A., Chen, W., Rasmussen, K. D., Zhu, X., Lundhaug, M., Müller, D. B., ... & Liu, G. (2022). Battery technology and recycling alone will not save the electric mobility transition from future cobalt shortages. *Nature communications*, *13*(1), 1341.

Zimmer, F., Scheibe, K., Stock, M., & Stock, W. G. (2019). Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice*, *7*(2), 40–53.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. Profile Books.

Zuiderveen Borgesius, F. J., Moeller, J., Kruikemeier, S., Fathaigh, R., Irion, K., Dobber, T., Bodó, B., & de Vreese, C. H. (2018). Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review*, *14*(1), 82–89.

Viranomaislähteet:

Eduskunnan oikeusasiamies. (2020). Eduskunnan oikeusasiamiehen kertomus vuodelta 2019. OAK 15/2020. Saatavilla <https://www.oikeusasiamies.fi/ documents/20184/42383/2019-fi/51758de7-f75b-449c-8967-a5372e40df0b>.

Euroopan komissio. (2018.) Tekoäly Euroopassa. Saatavilla <https://eur-lex.europa.eu/legal-content/FI/TXT/PDF/?uri=CELEX:52018DC0237&from=EN>

Euroopan komissio. (2020). Valkoinen kirja tekoälystä. Saatavilla <https://eur-lex.europa.eu/legal-content/FI/TXT/?uri=celex%3A52020DC0065>

Euroopan komissio. 2021/0106 (COD). Proposal for Artificial Intelligence Act. Saatavilla <https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-COM-Proposal-21-April-21.pdf>

Euroopan komissio. COM/2022/496 final. Proposal for AI Liability Directive. Saatavilla <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022PC0496>

Euroopan parlamentti. P9_TA(2023)0236. Artificial Intelligence Act. Amendments adopted by the European Parliament. Saatavilla <https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf>

Eurooppa-neuvosto. General approach on EU Artificial Intelligence Act. (14954/22). Saatavilla <https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-CZ-%E2%80%93-General-Approach-25-Nov-22.pdf>

EU-tuomioistuin, tapaus C-511/18: La Quadrature du Net. Saatavilla <https://curia.europa.eu/juris/document/document.jsf?text=&docid=232084&pageIndex=0&doclang=FI&mode=lst&dir=&occ=first&part=1&cid=229972>

Hallituksen turvapaikkapoliittinen toimenpideohjelma. (2015). Saatavilla <https://valtioneuvosto.fi/documents/10184/1058456/Hallituksen+turvapaikkapoliittinen+toimenpideohjelma+8.12.2015/98990892-c08e-4891-8c23-0d229f1d6099>.

Hallitusohjelma 2023. Vahva ja välittävä Suomi. Saatavilla <https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/165042/Paaministeri-Petteri-Orpon-hallituksen-ohjelma-20062023.pdf?sequence=1&isAllowed=y>.

Hallitusohjelma 2019. Osallistava ja osaava Suomi – sosiaalisesti, taloudellisesti ja ekologisesti kestävä yhteiskunta. Saatavilla <https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161931/VN_2019_31.pdf?sequence=1&isAllowed=y>.

Hallitusohjelma 2015. Ratkaisujen Suomi. Saatavilla <https://valtioneuvosto.fi/documents/10184/1427398/Ratkaisujen+Suomi_FI_YHDISTETTY_netti.pdf>.

HE 145/2022. Hallituksen esitys eduskunnalle julkisen hallinnon automaattista päätöksentekoa koskevaksi lainsäädännöksi.

HE 18/2019. Hallituksen esitys eduskunnalle laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

HE 224/2018. Hallituksen esitys eduskunnalle laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

Oikeuskanslerin virasto. (2020). Valtioneuvoston oikeuskanslerin kertomus vuodelta 2019. OKK 12/2020. Saatavilla <https://www.eduskunta.fi/FI/vaski/Kertomus/Documents/K_12+2020.pdf>.

PeVL 7/2019. Perustusvaliokunnan lausunto hallituksen esityksestä laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

PeVL 62/2018. Perustusvaliokunnan lausunto hallituksen esityksestä laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi.

Sosiaali- ja terveysministeriö. (2023). Työmarkkinatuen ja toimeentulotuen uudistuksen vaikutukset. Hallitusneuvotteluihin annettu lausunto. Saatavilla <https://stm.fi/documents/1271139/165490495/Tietopyynt%C3%B6-Ty%C3%B6markkinatuen+ja+toimeentulotuen+uudistuksen+vaikutukset.pdf>.

Tekoälyä käsittelevä korkean tason asiantuntijaryhmä. (2019). *Luotettavaa tekoälyä koskevat eettiset ohjeet.*

Työ- ja elinkeinoministeriö. (2022) *Tekoäly 4.0 -ohjelman loppuraportti*. Työ- ja elinkeinoministeriön julkaisuja 2022:60.

Työ- ja elinkeinoministeriö. (2019*). Edelläkävijänä tekoälyaikaan: Tekoälyohjelman loppuraportti.* Työ- ja elinkeinoministeriön julkaisuja 2019:23.

Työ- ja elinkeinoministeriö. (2017). *Suomen tekoälyaika: Suomi tekoälyn soveltamisen kärkimaaksi: Tavoite ja toimenpidesuositukset.* Työ- ja elinkeinoministeriön julkaisuja 41/2017.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2024

ISBN 978-951-29-9824-1 (painettu)
ISBN 978-951-29-9825-8 (verkko)
ISSN 0082-6987 (painettu)
ISSN 2343-3191 (verkko)