Mika Viljanen

Menikö juna jo? Tekoälyn sääntelemisen mahdollisuuksista

Lakimies
7–8/2023
s. 1204–1231

HAKUSANAT: tekoäly, sääntely, sääntelymahdollisuudet

1. Johdanto

Tekoälystä on kohistu viiden viime vuoden aikana paljon. Reilussa vuosikymmenessä toiveiden ja huolenaiheiden kehykset ovat liikkuneet Big Datasta algoritmeihin, koneoppimiseen ja lopulta generatiiviseen tekoälyyn. Kehitys on ollut nopeaa. Tekoälysovellukset näyttävät nyt selviävän vaativista kognitiivisista tehtävistä ja hallitsevan valtavia tietoaineistoja, ja ne ovat tulossa laajalti saataville.

Politiikka ja sääntely ovat seuranneet teknologisen kehityksen perässä, mutta hitaasti. Hallitukset ja muut politiikkatoimijat ovat laatineet tekoälystrategioita[1], -periaatteita ja -suuntalinjoja[2]. Ensimmäiset sääntelyhankkeet ovat myös käynnissä. Kovin pitkällä ei kuitenkaan olla globaalisti, sillä EU:n tekoälysäädös[3] ja tekoälyvastuudirektiivi[4] lienevät hankkeista merkittävimmät. Niidenkään yksi-

* *Mika Viljanen*, OTT, professori, Turun yliopisto. Kirjoitus on laadittu osana Suomen Akatemian rahoittamaa ETAIROS-projektia (päätösnumerot 327357 ja 352445).

1. Ks. kokoavasti Christian Djeffal – Markus B. Siewert – Stefan Wurster, Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies. Journal of European Public Policy 29(11) 2022, s. 1799–1821.
2. Jo vuonna 2019 Anna Jobin, Marcello Ienca ja Effy Vayena (The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9) 2019, s. 389–399) luettelivat kymmeniä periaate- ja suuntalinjadokumentteja. Ks. myös Graeme Auld – Ashley Casovan – Amanda Clarke – Benjamin Faveri, Governing AI through ethical standards: learning from the experiences of other private governance initiatives. Journal of European Public Policy 29(11) 2022, s. 1822–1844.
3. Ks. Ehdotus Euroopan parlamentin ja neuvoston asetukseksi tekoälyä koskevista yhdenmukaistetuista säännöistä (tekoälysäädös) ja tiettyjen unionin säädösten muuttamisesta. COM(2021) 206 final. Bryssel, 21.4.2021; Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach. Brussels 25 November 2022 ja Euroopan parlamentin tarkistukset 14. kesäkuuta 2023 ehdotukseen Euroopan parlamentin ja neuvoston asetukseksi tekoälyä koskevista yhdenmukaistetuista säännöistä (tekoälysäädös) ja tiettyjen unionin säädösten muuttamisesta (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).
4. Ehdotus Euroopan parlamentin ja neuvoston direktiiviksi sopimuksenulkoista siviilioikeudellista vastuuta koskevien sääntöjen mukauttamisesta tekoälyyn (direktiivi tekoälyyn liittyvästä vastuusta). COM(2022) 496 final. Bryssel, 28.9.2022.

tyiskohdista ei ole lokakuun 2023 alussa vielä selvyyttä. Kotimaisesta sääntelystä on vain pistemäinen esimerkki: maaliskuussa 2023 säädettiin yleisistä edellytyksistä automaattisten päätöstentekojärjestelmien käyttämiselle julkishallinnossa.[5]

Vaikka sääntelyhankkeet ovat EU:ta lukuun ottamatta vielä piirustuspöydillä, tekoälyn sääntely- ja yhteiskunnallisten vaikutusten hallintatavoista[6] ja oikeudellisesta sääntelystä[7] on jo kirjoitettu paljon. Keskustelussa tekoäly-ymmärrys on vaihdellut paljon. Yksinkertaisimmillaan on puhuttu musta laatikko -algoritmeista ja niiden hallinnasta.[8] Pahimmillaan tekoäly on ymmärretty vahvan autonomiseksi, ennakoimattomaksi, käsittämättömäksi ja oppivaksi olennoksi, jonka toimintatavoista ei voida saada selvyyttä ja joka vielä koko ajan muuttuu.[9]

Politiikkasuositukset ovat vaihdelleet. On esitetty esimerkiksi, että tärkeää on hallita erilaisia vinoumia hallitsemalla dataa[10], varmistaa järjestelmien läpinäkyvyys[11] ja selitettävyys[12], ja murehdittu sitä, että kehittämiseen ja käyttämiseen

5. Laki hallintolain muuttamisesta (487/2023). Ks. myös hallituksen esitys eduskunnalle julkisen hallinnon automaattista päätöksentekoa koskevaksi lainsäädännöksi 145/2022 vp.
6. Kirjoista esim. Virginia Dignum, Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer 2019; Justin Bullock (ed.), The Oxford handbook of AI governance. Oxford University Press 2022 ja Andrej Zwitter – Oskar Gstrein (eds), Handbook on the politics and governance of Big Data and artificial Intelligence. Edward Elgar 2023.
7. Kirjoista esim. Jacob Turner, Robot rules: regulating artificial intelligence. Springer 2019; Thomas Wischmeyer – Timo Rademacher (eds), Regulating Artificial Intelligence. Springer International Publishing 2020; Simon Chesterman, We, the robots? Cambridge University Press 2021 ja Charles Kerrigan (ed.), Artificial intelligence: law and regulation. Edward Elgar 2022.
8. Ks. esim. Frank Pasquale, The black box society: the secret algorithms that control money and information. Harvard University Press paperback edition. Harvard University Press 2016.
9. Tekoälyn ominaisuudet jäävät usein yksityiskohtaisesti määrittelemättä tai määritelmiin jätetään liikkumavaraa. Määritelmät ”vuotavat” ylöspäin. Niihin otetaan mahdollisuus siitä, että tulevaisuudessa tekoälystä voi tulla kognitiivisesti ihmiseen vertautuva oppiva olento. Ks. esim. Samir Chopra – Laurence F. White, A legal theory for autonomous artificial agents. University of Michigan Press 2011, s. 9–11; Visa Kurki, Voiko tekoäly olla oikeussubjekti? Lakimies 7–8/2018, s. 820–839, 821, 831–837 ja kriittisesti Neil M. Richards – William D. Smart, How should the law think about robots?, s. 3–22 teoksessa Ryan Calo – A. Michael Froomkin – Ian Kerr (eds), Robot law. Edward Elgar Publishing 2016, s. 20–21.
10. Ks. esim. Cathy O’Neil, Weapons of math destruction: how big data increases inequality and threatens democracy. Crown 2016.
11. Ks. esim. Heike Felzmann – Eduard Fosch Villaronga – Christoph Lutz – Aurelia Tamò-Larrieux, Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society 6(1) 2019: 2053951719860542.
12. Ks. esim. Andrew D. Selbst – Solon Barocas, The intuitive appeal of explainable machines. Fordham Law Review 87(3) 2018, s. 1085–1139; Daniel Vale – Ali El-Sharif – Muhammed Ali, Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI and Ethics 2(1) 2022, s. 815–826.

osalliset eivät ole vastuunalaisia.[13] Usein on myös taitettu peistä siitä, voivatko tekoälyjärjestelmät olla oikeushenkilöitä.[14]

Viimeistään vuoden 2023 alussa kävi kuitenkin selväksi, että sääntelykeskustelun tila on ongelmallinen. Teknologiaymmärrys ja sääntelyteemat ovat yhtä aikaa aikaansa perässä ja sitä edellä. Kun neuroverkoissa alkaa olla satoja kerroksia ja satoja miljardeja parametrejä kuten esimerkiksi ChatGPT:ssä[15], läpinäkyvyys ja selitettävyys sekä datalla ohjaaminen näyttävät hyödyttömiltä sääntelysuunnilta. Vaikka järjestelmän lähdekoodi olisi läpinäkyvää, kukaan ei ymmärtäisi sitä eikä neuroverkon toimintaa voitaisi mielekkäästi selittää. Koulutusdatan kuratointi on mahdotonta, kun järjestelmä on kehitetty käytännössä kaikella internetistä saatavalla tekstillä. Keinot varmistaa, että järjestelmät tuotokset ovat hyväksyttäviä, tai kohdistaa vastuu järjestelmän virheiden seurauksista mielekkäälle vastuutaholle ovat niin ikään vähissä. Toisaalta myös vinkeimmät kuvitelmat tekoälyn toimijuudesta ovat edelleen tieteiskirjallisuutta.

Tekoälysääntelykeskustelu onkin valinkauhassa. Tekoälystä on teknologiaryppäänä vaikea saada otetta, jotta sääntelyä voisi hahmotella. Usein kuvitellaan olemattomia ja rakennetaan sääntelyhahmotukset tieteiskirjallisuuden varaan. Toisaalta monet vakiintuneetkin ehdotukset näyttävät jääneen tekniikan kehityksestä. Siksi on syytä palata perusasioihin sääntelykentällä.

Pyrin tässä kirjoituksessa ensin sanoittamaan sitä, mitkä tekoälyteknologioiden tekniset ja sosiotekniset ominaisuudet – minkälainen tekoäly-ymmärrys – voivat olla sääntelysuunnittelun kannalta relevantteja. Tämän jälkeen pohdin, minkälaisilla sääntelykeinoilla voitaisiin vastata haasteisiin, jotka tästä tekoälymieliymmärryksestä nousevat. Kirjoituksesta piirtyy samalla käsitys tekoälysääntelyn mahdollisuuksista ja rajoista, kun hahmottelen eri ominaisuuskehysten avaamia ja sulkemia sääntelymahdollisuuksia.

Näkymä on toiveikkaan lohduton. *Mark A. Wood* on erottanut teknologiahaitoissa kaksi eri ulottuvuutta: teknologioilla voi olla joko välineellisiä tai generatiivisia haittavaikutuksia.[16] Välineelliset haittavaikutukset syntyvät, kun teknologiaa käytetään joko sen suunniteltuun tai odottamattomaan käyttötarkoitukseen. Generatiiviset haitat puolestaan syntyvät, kun teknologiat muut-

13. Ks. esim. Andreas Matthias, The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology 6(3) 2004, s. 175–183 ja Filippo Santoni de Sio – Giulio Mecacci, Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology 34(4) 2021, s. 1057–1084.
14. Ks. esim. David J. Gunkel, The other question: can and should robots have rights? Ethics and Information Technology 20(2) 2018, s. 87–99.
15. Partha Pratim Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 3 (2023), s. 121–154.
16. Ks. Mary Christina Wood, Atmospheric recovery litigation around the world: gaining natural resource damages against carbon majors to fund a sky cleanup for climate restoration. Research Handbook on Climate Change Law and Loss & Damage 2021, s. 512–513.

tavat toimintamahdollisuuksia. Autonomiset ajoneuvot ovat hyvä esimerkki hallinnan ongelmakentästä. Välineellisiä haittavaikutuksia on vielä mahdollista hallita sääntelyllä. Vaikka autonomisten ajoneuvojen tekoälykomponentit ovat monimutkaisia, tulkitsemattomia ja usein epälineaarisia, ajoneuvojen suunnitellun käytön ja väärinkäytönkin vaikutuksiin voidaan vaikuttaa esimerkiksi testaus- ja simulaatiosääntelyllä. Generatiivisissa haittavaikutuksissa tilanne on sen sijaan hankala. Tiedämme, että esimerkiksi autonomiset ajoneuvot voivat muuttaa fundamentaalisesti sitä, miten ja missä elämme, kenellä on työtä[17] tai miten paljon hiilidioksidia pääsee ilmakehään[18]. Realistisia keinoja ja poliittista tahtoa tällaisten makrotason kehityskulkujen oikeudelliseen hallintaan näyttäisi olevan vähän.[19]

Siksi keskityn siihen, missä toivoa on eli instrumentaalisten haittojen mikrotason hallintakeinoihin. Aloitan kirjoituksen johdattamalla lukijan toisessa luvussa siihen, miten tekoälyn ominaisuuksia olisi syytä hahmottaa ja sanoittaa, jotta relevantteja sääntelyongelmia voitaisiin tunnistaa ja niihin kehittää ratkaisuja. Seuraavissa kuudessa luvussa tarkastelen, millaisia sääntelyvaihtoehtoja tekoälyjärjestelmien eri teknisten ongelmien hallitsemiseksi on olemassa. Viimeisessä luvussa tiivistän käsitykseni tilannekuvasta.

2. Tekoäly ja tekoälyn tavanomaiset ongelmat

2.1. Mitä tekoäly on?

Tekoälyteknologioista on vaikea saada otetta. Erilaisia teknologioita on lukemattomia, ne ovat monimuotoisia, eikä niillä ole aina yhteisiä piirteitä.

Esimerkiksi *Russell* ja *Norvig* hahmottavat perusesityksessään tekoälyä nelikentällä. Tekoälyteknologiat voidaan erottaa toisistaan ensinnäkin sen perusteella, pidetäänkö älykkyyden merkkinä ihmisenkaltaista vai rationaalista toimintaa. Toiseksi teknologioita erottaa toisistaan se, onko äly prosesseja vai prosessien lopputuloksia. Näin järjestäen tekoälyä ovat teknologiat, jotka (1)

17. Ks. esim. Fábio Duarte – Carlo Ratti, The Impact of Autonomous Vehicles on Cities: A Review. Journal of Urban Technology 25(4) 2018, s. 3–18 ja David Bissell – Thomas Birtchnell – Anthony Elliott – Eric L Hsu, Autonomous automobilities: The social impacts of driverless vehicles. Current Sociology 68(1) 2020, s. 116–134.
18. Ks esim. Keigo Akimoto – Fuminori Sano – Junichiro Oda, Impacts of ride and car-sharing associated with fully autonomous cars on global energy consumptions and carbon dioxide emissions. Technological Forecasting and Social Change 174(6) 2022: 121311.
19. Ks. esim. Noam Kolt, Algorithmic Black Swans. Social Science Research Network 2023 osoitteessa https://papers.ssrn.com/abstract=4370566 (vierailtu 14.10.2023).

suorittavat kognitiivisia tehtäviä kuten ihmiset (ihmisprosessitekoäly), (2) tuottavat samoja kognitiivisia lopputuloksia kuin ihmiset (ihmistulostekoäly), (3) suorittavat kognitiivisia tehtäviä rationaalisilla tavoilla (rationaalisten prosessien tekoäly) tai (4) tuottavat rationaalisia lopputuloksia (rationaalisten lopputulosten tekoäly).[20]

Kaikissa määritelmissä on ongelmansa. Ihmiskehystyksessä emme aina tiedä, miten ihmiset toimivat tai mitä lopputuloksia ihmisten toiminnalla olisi. Vaikka tietäisimme, toimintatavat ja lopputulokset voivat olla epätoivottavia. Lisäksi koneet voivat joskus toimia ihmisiä paremmin ja suoriutua tehtävistä, joista ihmiset eivät selviä. Rationaalisuus on yhtä lailla pulmallinen kehys. Emme useinkaan varmasti tiedä, mitä rationaaliset toimintatavat tai lopputulokset olisivat. Rationaalisuuden kriteerit ovat normatiivisesti kiistanalaisia ja kulttuurisidonnaisia, kuten esimerkiksi feministiset tutkijat ovat lukemattomia kertoja esittäneet.[21] Joskus emme edes pysty ymmärtämään, miten koneet toimivat.[22] Tällöin koneiden toimintatavan rationaalisuutta ei voida arvioida. Tuloksellisuudenkin metriikatkin ovat väistämättä normatiivisia.

Tekoälyä voi hahmottaa myös teknologialähtöisesti. Nykyään tekoäly yhdistetään ensi sijassa koneoppimiseen (machine learning). Tässä kehyksessä tekoälyä ovat järjestelmät, jotka ovat syntyneet kone- ja syväoppimismenetelmillä.[23] Jos vain koneoppimisjärjestelmiä pidetään tekoälynä, vielä 2000-luvun edistykselliset tilastolliset Big Data -algoritmit[24] ja 1980- ja 1990-lukujen ”läpiohjelmoidut” perinteiset if, then -asiantuntijajärjestelmätkin jäävät katveeseen.[25] Samalla on hyvä pitää mielessä, että erityisesti robotiikassa tekoälyjärjestelmät ovat hyvin

20. Stuart Russell – Peter Norvig, Artificial intelligence. A modern approach. 4th edition. Pearson 2020, s. 20–23.
21. Esimerkkejä on lukemattomia Donna Harawaysta alkaen: Donna Haraway, Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies 14(3) 1988, s. 575–599; kiinnostavasta ns. queering-luennasta Blair Attard-Frost, Queering intelligence: A theory of intelligence as performance and a critique of individual and artificial intelligence, s. 23–39 teoksessa Michael Klipphahn-Karge – Ann-Kathrin Koster – Sara Morais dos Santos Bruss (eds), Queer reflections on AI. Routledge 2023.
22. Näin voi olla esimerkiksi valvomattomassa oppimisessa. Koneoppimisella voi syntyä päätösperusteita, jotka eivät merkitse ihmisille mitään mutta silti ehkä erottavat merkityksellisiä seikkoja. Ks. esim. Louise Amoore, Cloud ethics: algorithms and the attributes of ourselves and others. Duke University Press 2020 ja Justin Joque, Revolutionary mathematics: artificial intelligence, statistics, and the logic of capitalism. Verso 2022.
23. Ks. esim. Lior Rokach – Oded Maimon – Erez Shmueli (eds), Machine learning for data science handbook: data mining and knowledge discovery handbook. Springer 2023.
24. Ks. esim. Solon Barocas – Andrew D. Selbst, Big Data’s disparate impact. California Law Review 104(3) 2016, s. 671–732 ja historiasta Francis X. Diebold, On the origin(s) and development of the term “Big Data.” Social Science Research Network 2012 osoitteessa https://papers.ssrn.com/abstract=2152421 (vierailtu 5.5.2023).
25. Asiantuntijajärjestelmistä ks. Nils J. Nilsson, The Quest for artificial intelligence. Cambridge University Press 2009, luku 18.

monimutkaisia kokonaisuuksia, joissa lukuisat algoritmit toimivat ja käsittelevät tietoa rinnakkain ja peräkkäin.

Oikeaoppisen tekoälyn määritelmän jäljillä voidaankin juosta loputtomiin. Tässä kirjoituksessa suhtaudun pragmaattisesti kysymykseen siitä, mitä tekoäly on. Määritelmä on laaja. Tekoälyä ovat kaikki sosiotekniset järjestelmät, joissa on ei-biologisia elektronisia laskentakomponentteja ja joilla tietoa käsitellään ja tuotetaan erilaisia tuotoksia. Voiko tekoälyä säännellä tehokkaasti, riippuu siitä, minkälaisia järjestelmät ominaisuuksiltaan ovat.

2.2. Tekoälyn tavanomaiset ominaisuudet ja sääntelytavat

Kun tekoälyn sääntelystä keskustellaan tai sitä suunnitellaan, tekoäly täytyy teknologiana sanoittaa. Sanoitusvaihtoehtoja on esitetty keskustelussa lukemattomia, mutta eräät sanat toistuvat usein.[26]

Ennakoimattomuus on ehkä yleisin ominaisuus, joka tekoälyjärjestelmille osoitetaan: järjestelmät saattavat toimia tavoilla, joita niiden suunnittelijat ja käyttäjät eivät pysty ennakoimaan. Niillä on niin sanottuja emergenttejä ominaisuuksia.[27] Ennakoitavuuden teema liittyy läheisesti kolmeen muuhun huolenaiheeseen. Järjestelmien toimintaa ei aina jäännöksettä pystytä selittämään, koska ne jäävät läpinäkymättömiksi[28] mustiksi laatikoiksi[29], koska niiden sisälle ei eri syistä päästä katsomaan. Niitä ei toiseksi voida aina ymmärtää täysin:

26. Kirjallisuuskatsauksista ks. esim. Joseph E. Borson – Huan Xu, A path dependent approach for characterizing the legal governance of autonomous systems. IEEE Access 10 (2022): 119985–119998; Anja Folberth – Jutta Jahnel – Jascha Bareis – Carsten Orwat – Christian Wadephul, Tackling problems, harvesting benefits – A systematic review of the regulatory debate around AI. arXiv.org 2022 osoitteessa http://arxiv.org/abs/2209.05468 (vierailtu 2.9.2023) ja Patricia Almeida – Carlos Santos – Josivania Silva Farias, Artificial intelligence regulation: A meta-framework for formulation and governance. Proceedings of the 53rd Hawaii International Conference on System Sciences 2020 osoitteessa http://scholarspace.manoa.hawaii.edu/handle/10125/64389 (vierailtu 22.2.2021).
27. Ks. esim. Matthew U. Scherer, Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harvard Journal of Law & Technology 29(2) 2015, s. 353–400 ja Ryan Calo, Robotics and the lessons of cyberlaw. California Law Review 103(3) 2015, s. 513–564, 538–540.
28. Ks. esim. Jenna Burrell, How the machine "thinks": understanding opacity in machine learning algorithms. Big Data & Society 3(1) 2016: 2053951715622512 ja Scherer 2015.
29. Ks. Pasquale 2016.

ne ovat episteemisesti selittämättömiä (unexplainable)[30], läpitunkemattomia (inscrutable)[31] tai outoja (strange)[32].

Koneoppimismenetelmien läpimurron jälkeen on hahmotettu, että tekoälyjärjestelmät oppivat. Kuvitellaan usein, että järjestelmät muuttuvat käytön aikana tai jopa kehittävät uusia järjestelmiä. Tällainen oppiva järjestelmä on epävakaa. Sen algoritmit voivat muuttua ilman, että muutoksia testattaisiin ja validoitaisiin ennen kuin ne päätyvät tuotannon aikaisiin järjestelmiin.[33] Tien päässä häämöttää superäly.[34] Joskus esitetään myös, että tekoälyjärjestelmät olisivat luonteeltaan vahvasti tai heikosti[35] autonomisia.[36] Vahvasti autonominen tekoäly pystyy määrittelemään itse toimintansa tavoitteet ja moraaliset lähtökohdat.[37] Heikon autonomiselta tekoälyltä odotetaan vähemmän, eräänlaista tehtäväautonomisuutta.[38] Tällöin järjestelmä pystyy toteuttamaan saamansa tehtävän ilman ihmisen välitöntä tukea tai ohjausta.

Näistä tekoälyn kuvitelluista ominaisuuksista kasvaa esille tavanomainen sääntelypaletti. Ennakoimattomuutta on jotenkin hallittava. Jos emme voi olla varmoja, miten kone toimii, sääntelyssä joudutaan etsimään tapoja, joilla koneista voidaan tehdä ymmärrettäviä. Strategiat kohdistuvat usein suoraan järjestelmien tekniseen kokoonpanoon: järjestelmät pyritään avaamaan tarkastelulle ja niiden algoritmit tekemään ihmisille ymmärrettäväksi.[39] Välimallin

30. Ks. Nathan Colaner, Is explainable artificial intelligence intrinsically valuable? AI & Society 37(1) 2022, s. 231–238 ja Selbst – Barocas 2018.
31. Ks. Brent Daniel Mittelstadt – Patrick Allo – Mariarosaria Taddeo – Sandra Wachter – Luciano Floridi, The ethics of algorithms: mapping the debate. Big Data & Society 3(2) 2016, s. 1–21.
32. Ks. Bartosz Brożek – Michał Furman – Marek Jakubiec – Bartłomiej Kucharzyk, The black box problem revisited. Real and imaginary challenges for automated legal decision making. Artificial Intelligence and Law 2023 osoitteessa https://doi.org/10.1007/s10506-023-09356-9 (vierailtu 5.5.2023).
33. Tässä on syytä huomata, että nyt tarkoitan oppimisella sitä, että järjestelmät on tuotettu koneoppimismenetelmillä, kuten on tavanomaista esimerkiksi tietojärjestelmätieteen kielenkäytössä. Ks. juristien edellä tarkoitetuista, usein häilyvistä, oppimiskuvitelmista esim. Turner 2019, s. 75–80; Woodrow Barfield, Towards a law of artificial intelligence, s. 2–39 teoksessa Woodrow Barfield – Ugo Pagallo (eds), Research handbook on the law of artificial intelligence. Edward Elgar Publishing 2018, s. 4–5 ja Erik Røsag, Diabolus ex machina: When an autonomous ship does the unexpected, s. 124–143 teoksessa Henrik Ringbom – Erik Røsæg – Trond Solvang (eds), Autonomous ships and the law. Routledge 2020, s. 129.
34. Ks. Nick Bostrom, Superintelligence. Paths, dangers, strategies. Oxford University Press 2014.
35. Erottelusta ks. Willem F. G. Haselager, Robotics, philosophy and the problems of autonomy. Pragmatics & Cognition 13(3) 2005, s. 515–532.
36. Ks. esim. Scherer 2015.
37. Ks. esim. Curtis E. A. Karnow, The application of traditional tort theory to embodied machine intelligence, s. 51–77 teoksessa Ryan Calo – A. Michael Froomkin – Ian Kerr (eds), Robot law. Edward Elgar Publishing 2016.
38. Ks. Dignum 2019, s. 18–20.
39. Sääntelytapa näkyy myös tekoälysäädösehdotuksen 13 ja 52 artiklan läpinäkyvyysvaatimuksissa sekä hallintolain (434/2003) uudessa 53 e §:ssä. Kirjallisuudesta ks. esim. Selbst – Barocas 2018; Hans de Bruijn – Martijn Warnier – Marijn Janssen, The perils and pitfalls of explainable AI:

vaihtoehdossa datanhallinnalla ja -kuratoinnilla voidaan vaikuttaa siihen, mitä järjestelmät koulutuksessa oppivat datasta.[40] Epäsuorat sääntelystrategiat kohdistuvat tavallisesti taas suunnittelun metaprosesseihin. Esimerkiksi tekoälysäädösehdotuksen 9 artiklan säännöksillä pyrittäisiin varmistamaan, että kehittäjät tunnistavat järjestelmiensä riskit ja minimoivat ne. Samoin paljon on puhuttu sidosryhmien osallistamisesta ja kehittäjäkunnan diversiteetin tärkeydestä.

Itsestään muuttuvien ja oppivien järjestelmien edessä sääntelijät taas näyttävät olevan aseettomia. Jos oppivia järjestelmiä ei kielletä kuten esimerkiksi hallintolain 53 e §:ssä, suoria sääntelykeinoja ei oikeastaan ole. Jäljelle jäävät vain erilaiset vastuusääntelystrategiat, joita on tarjottu vaihtoehdoiksi myös ennakoimattomuuden hallintaan. Niillä yritetään vaikuttaa tekoälyjärjestelmiin määräämällä siitä, kuka tai mikä taho on vastuussa oppivasta järjestelmästä ja vaikuttaa siihen, minkälaisia järjestelmiä syntyy.[41] Vahvan autonomisissa järjestelmissä ollaan sääntelymahdollisuuksien oppivien järjestelmien tavoin viimeisellä rajalla. Työkalupakkiin näyttävät jäävän enää ihmisiin kohdistuvaan sääntelyyn vertautuvat keinot. Päädymme helposti pohtimaan sitä, voisiko tekoäly tokeentua ja millä edellytyksillä esimerkiksi rangaistuksen tai vahingonkorvausvelvollisuuden uhasta.[42]

Strategies for explaining algorithmic decision-making. Government Information Quarterly 39(2) 2022: 101666; Philipp Hacker – Jan-Hendrik Passoth, Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond, s. 343–373 teoksessa Andreas Holzinger – Randy Goebel – Ruth Fong – Taesup Moon – Klaus-Robert Müller – Wojciech Samek (eds), xxAI – Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18 2020, Vienna, Austria. Revised and Extended Papers. Springer 2022.

40. Esimerkiksi tekoälysäädösehdotuksen 10 artiklassa edellytettäisiin, että data on virheetöntä, edustavaa ja kattavaa. Ks. myös Marijn Janssen – Paul Brous – Elsa Estevez – Luis S. Barbosa – Tomasz Janowski, Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly 37(3) 2020: 101493 ja Matti Mäntymäki – Matti Minkkinen – Teemu Birkstedt – Mika Viljanen, Defining organizational AI governance. AI and Ethics 2(4) 2022, s. 603–609.
41. Suomalaisista vastuusääntelyä koskevista puheenvuoroista ks. Felix Collin, Unmanned ships and fault as the basis of shipowner's liability, s. 85–97 teoksessa Autonomous Ships and the Law. Routledge 2020; Béatrice Schütte – Lotta Majewski – Katri Havu, Damages Liability for Harm Caused by Artificial Intelligence – EU Law in Flux. SSRN 2021 osoitteessa https://papers.ssrn.com/abstract=3897839 (vierailtu 24.1.2022) ja Lauri Luoto, Itsestään ajavat autot ja rikosoikeudellinen vastuu. Lakimies 6/2022, s. 927–948.
42. Ks. esim. Gabriel Hallevy, The Criminal Liability of Artificial Intelligence Entities – from Science Fiction to Legal Social Control. Akron Intellectual Property Journal 4(2) 2010, s. 171–201.

3. Tekoälysääntelyn uudet sanat

Tavanomaiset sanoitukset tuottavat vaillinaisen ja harhaanjohtavan kuvan tekoälyteknologioiden teknisistä haasteista ja sääntelyvaihtoehdoista. Jatkuvasti oppivia ja vahvan autonomisia järjestelmiä ei esimerkiksi juuri ole käytössä. Tekoälyjärjestelmät ovat osin ennakoimattomia ja ominaisuuksiltaan emergenttejä, mutta tosiasiassa järjestelmien toimintaa pystytään silti hallitsemaan muullakin kuin vain läpinäkyvyys-, selitettävyys- ja vastuusääntelyllä. Tavanomainen tekoäly-ymmärrys ohjaa sääntelykeskustelua vääriin suuntiin. Yhtäältä tehokkaita sääntelykeinoja ei huomata eikä niihin kiinnitetä riittävää huomiota. Toisaalta osa sääntelykeskustelusta on lähinnä tieteiskirjallisuutta. Jotta sääntely olisi mielekästä, sen täytyy kohdistua oikeisiin kohteisiin. Sitä varten tarvitaan oikeat sanoitukset, jotka ohjaavat sääntelyä oikeaan suuntaan.[43] Esitän seuraavassa oman hahmotukseni siitä, millaisilla sanoituksilla tekoälyjärjestelmistä ja niiden ominaisuuksista olisi syytä puhua, jotta sääntelyn haasteista ja mahdollisuuksista piirtyisi mielekäs kuva.

Esitän kuusi uutta sanoitusta, joilla puhua tekoälyn sääntelyn kannalta ongelmallisista teknisistä ulottuvuuksista. Ensimmäinen ulottuvuus tuo esille sen, että tekoälyteknologiat ovat sosio-teknisiä järjestelmiä. Ihmiset ovat niissä aina mukana, mutta ihmisten rooli ja käänteisesti järjestelmän kyky vaikuttaa ympäristöön itsenäisesti vaihtelevat (tekninen toimijuus). Toinen relevantti ulottuvuus on se, kuinka monimutkaisia järjestelmät ovat (monimutkaisuus). Kolmanneksi järjestelmät ja niiden toimintatavat ovat vaihtelevalla tavalla ihmisten tulkittavissa (tulkittavuus). Neljänneksi järjestelmien syöte- ja tuotosavaruuksien laajuus ja niiden toiminnan epälineaarisuuden (syötteiden ja tuotteiden moninaisuus) aste vaihtelevat. Viides ulottuvuus on se, toimivatko järjestelmät deterministisesti eli tuottavat aina saman tuloksen samoilla syötteillä (determinanssi). Kuudes ulottuvuus on se, pysyvätkö järjestelmän sisuskalut muuttumattomina käytön aikana vai muuttuvatko ne (dynaamisuus). Käsittelen seuraavassa kutakin ominaisuutta ja sen hallinnassa mahdollisia sääntelykeinoja.

43. Sanoitusten eli metaforien merkityksestä ks. Richards – Smart 2016, s. 16–18.

4. Tekninen toimijuus ja sen hallintakeinot

4.1. Tekninen toimijuus

Kun tekoälyjärjestelmiä tarkastellaan, on ensin tärkeää selvittää, minkälainen tekninen toimijuus on sillä sosioteknisellä järjestelmällä, jonka osa tekoälyjärjestelmä on, eli miten itsenäisesti järjestelmä pystyy aktuoimaan eli siirtämään tuotoksensa ympäristömuutoksiksi.[44] Matalan teknisen toimijuuden tekoälyjärjestelmissä tarvitaan ihmisen välittömiä interventioita, jotta järjestelmien tuotokset välittyisivät muutoksiksi ympäristössä. Korkean teknisen toimijuuden järjestelmä voi sen sijaan saada aikaan muutoksia ympäristössään ilman ihmisen myötävaikutusta. Välittömästi ihmisistä riippuvaisten ja autonomisten järjestelmien väliin jää laaja kirjo erilaisia sosioteknisiä sommitelmia, joissa ihmisen rooli koneen välittäjänä vaihtelee.

Korkea-asteinen tekninen toimijuus luo perustavan sääntelyongelman. Mitä itsenäisemmin järjestelmien toiminta välittyy ympäristöön, sitä todennäköisemmin perinteiset sääntelytavat menettävät otteensa.[45] Mekanismi on yksinkertainen. Perinteisesti sääntely on kohdistunut ihmisiin. Esimerkiksi rikoslaissa kielletään ihmisiä tappamasta toisiaan. Vahingonkorvausvelvollisuudella voidaan viestiä, että valintoja, joilla on tiettyjä seurauksia, olisi syytä välttää. *Fredrick Schauer* on kuvannut mekanismia esittämällä, että oikeus tuottaa ihmisille syitä toimia oikeuden edellyttämällä tavalla.[46] Oikeudella on materiaaliset edellytyksensä. Ihmisten kehojen, kognitioiden ja sosiaalisen elämän monimutkainen kokonaisuus tuottavat alustan, johon oikeus voi tarttua. Jos ihmisen kehot ja kognitio eivät ole läsnä, oikeus on tyhjiä sanoja paperilla. Tämä on tekoälysääntelyn keskeinen ongelma. Kun koneet täyttävät maailman, oikeuden tila tyhjenee. Siellä ei ole enää mitään, mihin oikeudella tarttua.

4.2. Ihmiset koneiden valvojiksi

Tilanne ei ole kuitenkaan täysin toivoton. Ihmisten katoamiseen voidaan reagoida. Ensimmäinen keino on lukita ihmiset järjestelmiin. Tällöin koneen toimijuuden päälle kerrostetaan ihmistoimijuuden kerros, joka varmistaa, että ihmisten oikeuden ote koneesta pitää, kuten esimerkiksi tekoälysäädösehdotuksen 12

44. Ks. esim. Davide Luigi Totaro, Machine or Robot? Thoughts on the Legal Notion of Autonomy in the Context of Self-Driving Vehicles and Intelligent Machines. European Business Law Review 34(1) 2023, s. 104.
45. Ks. Mika Viljanen, A cyborg turn in law? German Law Journal 18(5) 2017, s. 1277–1308.
46. Ks. Frederick F. Schauer, The force of law. Harvard University Press 2015.

artiklan human oversight -säännöksissä. Kirjallisuudessa on tunnistettu ainakin kaksi mahdollista valvontakonstellaatiota. Ensimmäisessä ihminen on osa päätöksentekokiertoa (human-in-the-loop). Ihminen hyväksyy järjestelmän päätöksen ennen kuin järjestelmä aktuoi sen. Toinen vaihtoehto on "on-the-loop" eli seuraa järjestelmän toimintaa valmiina puuttumaan siihen, jos tarvetta on. Koneet kuitenkin aktuoivat päätöksensä itsenäisesti, jos ihminen ei sitä estä.[47]

Ratkaisuilla on rajansa. Ihmisen ja tietokoneen vuorovaikutuksen tutkimuksessa on moneen kertaan osoitettu, että ihmisen toimijuus muuttuu, kun sosioteknisen järjestelmään lisätään teknologisia kerroksia. Muutokset eivät aina ole myönteisiä. Ihmisistä voi esimerkiksi tulla huolimattomia ja välinpitämättömiä, kun koneet tekevät päätökset, tai he unohtavat tai eivät koskaan opi kunnolla tehtäviään.[48] Siksi teknisen toimijuuden rajoittaminen ei aina auta, ja koneiden ja ihmisten ”tiimityön” järjestelyt on joka tapauksessa suunniteltava tarkoin.

4.3. Katse metatoimintaan

Valvontajärjestelyjen lisäksi toinen vaihtoehto on, että vastuusääntelyssä katse siirretään siihen inhimilliseen tausta- eli metatoimintaan, joka määrää, mitä koneet tekevät. Tällöin esimerkiksi rangaistus- ja vahingonkorvausvastuun säännöillä voitaisiin edelleenkin ohjata toimintaa, kun säännöt sovitetaan ihmisten välittömän toiminnan sijaan ohjaamaan suunnitteluprosesseja tai käyttöönottopäätöksiä.

Metaprosessien sääntäleminen vastuusäännöillä on kuitenkin ongelmallista. Vastuusääntelyssä on kehittynyt vuosisatainen kokemus siitä, miten muotoillaan käyttökelpoisia sääntöjä ohjaamaan inhimillisiä päätöksentekoprosesseja, joiden seuraukset ovat lähellä päätöksiä. Jos tarkoituksena on ohjata metatoimintaa, säännöt natisevat liitoksissaan.

Otetaan esimerkiksi tuottamus vahingonkorvausoikeudessa. Useimmissa länsimaisissa oikeusjärjestyksissä tuottamus on vahingonkorvausvastuun perusvastuuperuste. Vahingonaiheuttaja joutuu korvaamaan aiheuttamansa va-

47. Ks. ihmiskerroksista esim. Meg Leta Jones, The right to a human in the loop: Political constructions of computer automation and personhood. Social Studies of Science 47(2) 2017, s. 216–239.

48. Ks. esim. Victor Riley, Operator reliance on automation: theory and data, s. 19–35 teoksessa Automation and human performance: theory and applications. Lawrence Erlbaum 1996; Christopher D. Wickens – Benjamin A. Clegg – Alex Z. Vieane – Angelia L. Sebok, Complacency and automation bias in the use of imperfect automation. Human Factors 57(5) 2015, s. 728–739 ja R. Parasuraman – T.B. Sheridan – C.D. Wickens, A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 30(3) 2000, s. 286–297.

hingon, jos hän on aiheuttanut sen tuottamuksellaan eli väärällä valinnallaan.[49] Tekoälykonteksteissa tuottamusvastuusta tulee hankala sääntelyväline, ainakin jos tarkoituksena on vaikuttaa kehittäjien käyttäytymiseen. Tuottamuksen määrittely on yksityiskohdissaan arvionvaraista jo silloin, kun pohditaan, minkälaisia välittömiä riskejä ihmiset saavat ottaa. Kun siirrytään pohtimaan, minkälaiset suunnitteluvalinnat ovat huolellisia, vaikeuskerroin kasvaa ainakin kertaluokalla ja arvioinnista tulee hidasta jälkiviisastelua. Kun vahingonkorvausoikeudenkäynti päättyy vuosien prosessin jälkeen, käsillä on toimintanormi, joka on auttamattomasti kiinni yksittäistapauksen olosuhteissa ja koskee todennäköisesti teknologiaa, jota ei enää kehitetä.

Toinen ongelma juontuu siitä, että tekoälyteknologioiden metaprosesseissa päätökset hajautuvat sekä ajassa että paikassa. Joku ostaa tai kouluttaa algoritmin, jonka joku toinen integroi järjestelmään muiden algoritmien rinnalle. Tämän jälkeen järjestelmä testataan ja validoidaan, kunnes joku kolmas tekee päätöksen laskea järjestelmä markkinoille tai ottaa se käyttöön. Tällaiset hajaantuneet päätösprosessit ovat omiaan hankaloittamaan vastuun allokoimista erityisesti tuottamusvastuussa.[50] Yksikään yksilö tai edes organisaatio ei välttämättä hallitse koko metaprosessia, vaan järjestelmiin kerrostuu lukemattomia eriaikaisia ja -paikkaisia valintoja. Millään taholla ei välttämättä ole kokonaisnäkemystä siitä, minkälainen järjestelmä kokonaisuudessaan on tai miten järjestelmän eri osat vuorovaikuttavat. Lähestytään tilannetta, jossa kukaan tai mikään ei näissä oloissa voi enää valita väärin paitsi silloin, kun ottaa käyttöön järjestelmän, jonka riskejä ei voida enää tuntea.[51]

Lopputulos onkin, että jos tuottamusvastuun säännöt tuottavat ainoastaan epämääräisiä ohjeita eikä vääriä valintoja voida hevin tunnistaa, mahdollisuus ohjata suunnitteluprosesseja vastuusäännöillä kuihtuu pitkälti pois. Jäljelle jää ankara vastuu.[52] Ankarasta vastuusta säätämällä voidaan toki kannustaa esimer-

49. Ks. Juha Häyhä, Sopimusoikeus, vahingonkorvausoikeus ja väärät valinnat, s. 87–127 teoksessa Jukka-Pekka Takala – Kimmo Nuotio (toim.), Ymmärtäminen ja oikeudellinen vastuu. Edita 1997.
50. Ks. esim. Scherer 2015; Mark Coeckelbergh, Moral responsibility, technology, and experiences of the tragic: from Kierkegaard to offshore engineering. Science and Engineering Ethics 18(1) 2012, s. 35–48.
51. Ks. Mitja Kovac, Autonomous artificial intelligence and uncontemplated hazards: towards the optimal regulatory framework. European Journal of Risk Regulation 13(1) 2022, s. 94–113, 104–106.
52. Ks. S Li – M Faure – K Havu, Liability rules for AI-related harm: law and economics lessons for a European approach. European Journal of Risk Regulation 13(4) 2022, s. 618–634; Baris Soyer – Andrew Tettenborn, Artificial intelligence and civil liability – do we need a new regime? International Journal of Law and Information Technology 30(4) 2022, s. 385–397; Kovac 2022 ja Christiane Wendehorst, Liability for artificial intelligence: The need to address both safety risks and fundamental rights risks, s. 187–209 teoksessa Oliver Mueller – Philipp Kellmeyer – Silja Voeneky – Wolfram Burgard (eds), The Cambridge handbook of responsible artificial intelligence: interdisciplinary perspectives. Cambridge University Press 2022.

kiksi suunnittelijoita välttämään vahinkoja[53], mutta varsinaisia metatoiminnan käyttäytymisohjeita ne eivät tuota.

4.4. Koneet, jotka noudattavat oikeutta?

Jos valvonta- ja vastuusääntelyt jäävät vaatimattomiksi vaihtoehdoiksi tekoälyjärjestelmien kehityksen ohjaamisessa, teknisen toimijuuden hallintavaihtoehtoja jää kolme. Joko oikeus on ohjelmoitava koneisiin osaksi niiden päätöksentekojärjestelmiä, koneet on koulutettava käyttäytymään siten, että ne noudattavat oikeutta, tai koneisiin on rakennettava toiminnallisuuksia, jotka tuottavat niille kyvyn vaikuttua ihmisten oikeudesta.[54]

Konetta on vaikea ohjelmoida noudattamaan ihmisille tarkoitettuja sääntöjä. Kone tarvitsisi yksiselitteisiä, tarkasti määriteltyjä formaaleja sääntöjä, jotka toimisivat rajatuissa ontologioissa. Ihmisen säännöt eivät täytä näitä vaatimuksia. Jotta ohjelmoiminen olisi mahdollista, säännöt on formalisoitava eli käännettävä koneiden kielelle. Tehtävä ei ole triviaali.[55] Sitä voidaan kuitenkin helpottaa tekemällä oikeudesta helpompaa koneille. Niin sanotusta machine readable law -tutkimusta on jo jonkin verran.[56] Koneluettavan oikeuden ideaa on kuitenkin computational law -tutkimussuuntauksessa kritisoitu voimakkaasti. Esimerkiksi *Mireille Hildebrandt* on huomauttanut, että kun oikeus käännetään koneille, menetetään tärkeitä oikeuden ulottuvuuksia. Koneita varten oikeuden epävarmuudet ja epämääräisyydet on poistettava ja sen avoin, jatkuvasti päivittyvä tekstuuri on suljettava.[57]

53. Ks. esim. Andrew F. Daughety – Jennifer F. Reinganum, Economic analysis of products liability: theory, s. 69–95 teoksessa Jennifer H. Arlen (eds), Research handbook on the economics of torts. Edward Elgar Publishing 2013 ja Tim Friehe – Cat Lam Pham – Thomas J. Miceli, Product liability and strategic delegation: Endogenous manager incentives promote Strict Liability. Review of Industrial Organization 61(2) 2022, s. 149–169.
54. Ks. vaihtoehdoista autonomisessa liikenteessä esim. Henry Prakken, On the problem of making autonomous vehicles conform to traffic law. Artificial Intelligence and Law 25(3) 2017, s. 341–363.
55. Ks. esim. Hanif Bhuiyan – Guido Governatori – Andry Rakotonirainy – Meng Weng Wong – Avishkar Mahajan, Driving decision making of autonomous vehicle according to Queensland overtaking traffic rules. Springer Link 2023 osoitteessa https://doi.org/10.1007/s12626-023-00147-x (vierailtu 15.10.2023).
56. Ks. esim. Alice Witt – Anna Huggins – Guido Governatori – Joshua Buckley, Encoding legislation: a methodology for enhancing technical validation, legal alignment and interdisciplinarity. Artificial Intelligence and Law 2023 ja kokoavasti Christopher Markou – Simon Deakin, Ex machina lex: exploring the limits of legal computability, s. 31–66 teoksessa Simon Deakin – Christopher Markou (eds), Is law computable? Critical perspectives on law and artificial intelligence. Hart Publishing 2020.
57. Ks. Mireille Hildebrandt, Algorithmic regulation and the rule of law. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2128)

Toiseksi koneoppimismenetelmät ovat tehneet mahdolliseksi kuvitella, että koneet voisivat oppia oikeat toimintatavat koulutusdatasta. Lähestymistapa edellyttäisi tietysti massiivista määrää opetusskenaarioita mutta jättäisi epävarmaksi sen, toimisiko kone uudessa tilanteessa oikein.

Kolmas vaihtoehto on epämääräisin. On vaikea edes kuvitella, minkälaisia toiminnallisuuksia ja rakenteita koneet tarvitsisivat ymmärtääkseen oikeutta ja vaikuttuakseen välittömästi oikeudesta. Ainakin koneiden olisi kyettävä ymmärtämään luonnollista kieltä, täydentämään epämääräiset säännöt muulla oikeuslähdeaineksella ammattitaitoisten juristien tapaan, ennakoimaan vaikeissa tapauksissa, miten oikeus sulkeutuisi institutionaalisissa käytännöissä, hahmottamaan yleisten sääntöjen ja yksittäistapausten vuorovaikutusta, tunnistamaan ympäristösyötteistä objektit, jotka vastaavat sääntöjen tunnusmerkistöjä, ja sovittamaan ympäristön objektit eli faktapremissi yhteen normipremissin kanssa.[58] Lisäksi koneisiin täytyisi rakentaa toiminnallisuus, joka saisi ne välittämään siitä, mitä oikeus käskee niitä kussakin tapauksessa tekemään. Ehkä koneen voi ohjelmoida toteuttamaan oikeutta.[59] Ihmisistä ei vaikuta kuitenkaan olevan apua tässä ohjelmointiprosessissa. Emme tiedä, miksi ja miten oikeus ohjaa ihmisten toimintaa. Pikemminkin tiedämme ehkä sen, että oikeudella pääasiassa onnistutaan ohjaamaan ihmisten toimintaa sekä epävarmasti että tehottomasti.[60]

5. Monimutkaisuuden ongelmat

5.1. Monimutkaisia laitteita, monimutkaisia ongelmia

Seuraava haaste nousee monimutkaisuudesta. Tekoälyjärjestelmistä on tullut ja on tulossa jatkuvasti teknisesti aiempaa monimutkaisempia. Yhtäältä kysymys on siitä, että järjestelmillä pyritään hallitsemaan entistä hienosyisempiä prosesseja aiempaa tarkemmin ja yksityiskohtaisemmin. Tämä johti jo ennen koneoppimisvallankumousta siihen, että koodirivien määrä erilaisissa järjestel-

2018: 20170355; Mireille Hildebrandt, Code-driven law: Freezing the future and scaling the past, s. 67–83 teoksessa Simon Deakin – Christopher Markou (eds), Is Law Computable? Critical Perspectives on Law and Artificial Intelligence. Bloomsbury Publishing 2020 ja Sylvie Delacroix, Automated systems and the need for change, s. 161–175 teoksessa Simon Deakin – Christopher Markou (eds), Is law computable? Critical perspectives on law and artificial intelligence. Hart 2020.

58. Ks. kokoavasti esim. Markou – Deakin 2020.
59. Kurki (2018) pohtii, miten koneen saa tottelemaan sääntöjä, ja päätyy rankaisemiseen. Kone on ohjelmoitava välttämään rangaistusta.
60. Ks. Lawrence M Friedman, Impact. How law affects behavior. Harvard University Press 2016.

missä kasvoi räjähdysmäisesti.[61] Koneoppimisjärjestelmissä monimutkaisuuden ongelma nousee kokonaan uudelle tasolle. Neuroverkkojen koulutusdatamassat ovat kasvaneet ja tavoitetilat moninaistuneet, minkä lisäksi verkkojen kerrosten, kytkentöjen ja parametrien määrät ovat räjähtäneet käsiin. Esimerkiksi jo ChatGPT tuotti loputtoman kirjon erilaisia tekstikatkelmia ja neuroverkossa oli satakunta kerrosta ja satoja miljardeja parametrejä.[62] Onkin syytä huomata, että vaikka tekoälyjärjestelmien läpinäkymättömyys tai vaikeudet tulkita niitä esitetään usein syyksi sille, että järjestelmiä on vaikea säännellä, jo monimutkaisuus tekee järjestelmien toiminnan hallinnasta haastavaa.

Monimutkaisuus syventää vastuusääntelyn ongelmia. Käyttäjien ja kehittäjien kyky arvioida järjestelmiä heikkenee entisestään. Myös syy-seuraussuhteita yksittäisten suunnitteluvalintojen ja konkreettisten seurausten välillä voi olla vaikea hahmottaa. Monimutkaisuus johtaa kuitenkin myös toiseen, edellistä tärkeämpään suuntaan. Kun monimutkaisissa ohjelmistojärjestelmissä samaan toiminnallisuuteen voidaan päätyä usealla erilaisella teknisellä ratkaisulla, tavanomaisen teknisen käskysääntelyn rajat tulevat vastaan. Käy mahdottomaksi määrätä, mitä teknistä ratkaisua on käytettävä[63], kun selkeitä teknisiä reunaehtoja, joiden varassa rakentaa toimiva järjestelmä, ei enää ole. Jos ja kun halutaan lisäksi tukea innovaatioita, ei ole järkevää lukita toimialaa johonkin tiettyyn tekniseen ratkaisuun. Tarvitaan muita lähestymistapoja.

Täysin tuntemattomalla alueella ei jouduta kuitenkaan liikkumaan. Sääntelylle on yleistymässä olevia esikuvia esimerkiksi ajoneuvojen ja lääkkeiden sääntelykehyksissä, standardisointiorganisaatioiden prosessistandardeissa ja finanssitoimialalla. Teknisessä sääntelyssä suunta näyttääkin väistämättä kulkevan teknisestä käskysääntelystä kohti tulosperusteista suorituskykysääntelyä ja liikkeenjohdollista eli prosessisääntelyä. Käsittelen vaihtoehtoja seuraavaksi.

5.2. Suorituskykysääntely

Tulos- tai suorituskykyperusteisessa sääntelyssä fokus on säänneltävän prosessin lopputuloksessa. Sääntelyvaatimukset määräävät, minkälaisia prosessin lopputulokset saavat olla[64], mutta jättävät säänneltävien tehtäväksi valita, mitä tekniik-

61. Robert N Charette, This car runs on code. IEEE Spectrum 2009 osoitteessa https://spectrum.ieee.org/this-car-runs-on-code (vierailtu 1.9.2023).
62. Ks. esim. Ray 2023.
63. Ks. teknisten standardien historiasta JoAnne Yates – Craig Murphy, Engineering rules: global standard setting since 1880. Johns Hopkins University Press 2019 ja EU:n tuoteturvallisuussäänteleystä Jukka Ruohonen, A review of product safety regulations in the European Union. International Cybersecurity Law Review 3(2) 2022, s. 345–366.
64. Ks. Tobias D. Krafft – Katharina A. Zweig – Pascal D. König, How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. Regulation

kaa tulee käyttää tavoitteiden saavuttamiseksi. Tulos- ja suorituskykyperusteinen sääntely on sinänsä teknisen sääntelyn peruskauraa. Esimerkiksi ympäristöluvituksen päästörajoitteet ovat oppikirjaesimerkkejä tulos- tai suorituskykyperusteisesta sääntelystä.[65]

AI-konteksteissa tulos- ja suorituskykyperusteinen sääntely vaikuttaisi jäävän usein ainoaksi mielekkääksi sääntelyvaihtoehdoksi. Jos järjestelmät ovat monimutkaisia eikä niiden toimintaperiaatteita voida verifioida formaaleilla menetelmillä eli osoittamalla teoreettisesti, että ne toimivat oikein, järjestelmiä on testattava, jotta niiden tuottamia tuloksia voidaan tarkastella.[66] Sääntelykäytössä testaamisvaatimukset edellyttävät kuitenkin sitä, että sääntelijät onnistuvat artikuloimaan, minkälaiset testit ovat riittäviä. Lisäksi on määritettävä soveliaat metriikat järjestelmien hyväksyttävälle toiminnalle ja varmistuttava siitä, että järjestelmät testataan uskottavissa testausympäristöissä. Sääntelytapa vaatii siis merkittäviä resursseja ja asiantuntemusta, jos sääntelijät eivät turvaudu toimialan itsesääntelyyn. Palaan teemaan jäljempänä luvussa 7.2.

5.3. Prosessit kuntoon

Jälkimmäinen strategianippu eli prosessi- ja liikkeenjohdollinen sääntely on yleistynyt viime vuosikymmeninä liiketoiminta- ja teknologiasääntelyn eri konteksteissa. Sääntelytavassa pyritään muokkaamaan yritysten toimintaprosesseja siten, että tarkoituksenmukaisten suunnittelupäätösten todennäköisyys nousee ja epätoivottavista suunnitteluratkaisuista tulee harvinaisia. Käytännössä liikkeenjohdollinen ja prosessisääntely voi toimia esimerkiksi siten, että toimijat ohjataan tuottamaan tietoa soveliaista liiketoiminta- tai teknologiakehitysprosessien riskeistä tai seurauksista sekä laatimaan ja toteuttamaan erilaisia toimintasuunnitelmia niiden erilaisten riskien tai haittojen hallitsemiseksi.

EU:n tekoälysäädöksen 9 artikla on erinomainen esimerkki prosessisääntelystä. Artiklassa vaadittaisiin, että kaikilla tekoälykehittäjillä on riskienhallintajärjestelmä, jota käytetään, kun tekoälyjärjestelmiä kehitetään. Riskienhallintajärjestelmä on monimutkainen työtehtävien, organisaatioprosessien ja kykyjen kokonaisuus, joka varmistaa, että kehittäjäorganisaatio tunnistaa riskejä, joita sen kehittämästä tekoälyjärjestelmästä voi aiheutua, ja toimii tavoitteellisesti

& Governance 16(1) 2022, s. 119–136, 128–129.

65. Ks. yleisesti esim. Cary Coglianese – Jennifer Nash, The law of the test: performance-based regulation and diesel emissions control. Yale Journal on Regulation 34(1) 2017, s. 33–90.

66. Ks. esim. Deven R. Desai – Joshua A. Kroll, Trust but verify: a guide to algorithms and the law. Harvard Journal of Law & Technology 31(1) 2017–2018, s. 1–64 ja Mika Viljanen, Safety by simulation: theorizing the future of robot regulation. AI & Society 2023 osoitteessa https://doi.org/10.1007/s00146-023-01730-0 (vierailtu 17.10.2023).

näiden riskien poistamiseksi, vähentämiseksi, rajoittamiseksi ja hallitsemiseksi.[67] Kuinka tehokasta tällainen sääntely voi olla, jää nähtäväksi.

6. Tulkinnan vaikeudet

6.1. Mystisiä koneita

Tekoälyteknologioiden kolmas ongelma on, että niiden toimintaperiaatteita ei aina voida selittää, ymmärtää, tarkastella tai tulkita. Niin on sanottu lukemattomissa kommenteissa.[68] Selitettävyyden ja tulkinnat teemat nousivat AI-keskustelussa esille erityisesti sen jälkeen, kun Big Data -menetelmien kehitys lähti käyntiin 2000-luvulle tultaessa. Viimeistään syväoppimisen myötä selittämättömyydestä tuli yksi tekoälykeskustelun keskipisteistä. Taitavimmatkaan datatieteilijät eivät enää kyenneet täysin ymmärtämään niitä syötteiden transformaatiopolkuja, joiden tuloksena monimutkaiset neuroverkot päätyivät niihin lopputuloksiin, joihin päätyivät. Koneoppivista tekoälyjärjestelmistä oli tullut tulkitsemattomia (uninterpretable).[69] Ne eivät enää toimineet ihmisille tutuissa narratiivisissa ontologioissa eivätkä symbolisen logiikan varassa.[70] Jos menetelmät ovat tulkitsemattomia, ihmiset eivät voi simuloida niitä mielessään eivätkä kääntää niitä esimerkiksi loogisiksi jos, sitten -lausekokonaisuuksiksi.

Jos tekoälyn toimintaa ei voi hahmottaa ja kääntää ihmisille ymmärrettävälle kielelle ja logiikalle, joudutaan vaikeuksiin kahdessa suunnassa. Ensimmäinen suunta tiivistyy selitettävyysvaatimuksiin. Toinen jatkaa teknisen käskysääntelyn ahdinkoa.

67. Sääntelytavan juurista ks. Yates – Murphy 2019, luku 9 ja Christine Parker, Meta-regulation: legal accountability for corporate social responsibility, s. 335–368 teoksessa David Kinley (ed.), Human rights and corporations. Routledge 2009.
68. Ks. esim. Mittelstadt – Allo – Taddeo – Wachter – Floridi 2016 ja Selbst – Barocas 2018.
69. Ks. esim. Zachary C. Lipton, The mythos of model interpretability. Queue 16(3) 2018, s. 31–57 ja Ayush Somani – Alexander Horsch – Dilip K Prasad, Interpretability in deep learning. Springer 2023.
70. Ks. esim. Burrell 2016 ja Sean Gerrish, How smart machines think. The MIT Press 2018.

6.2. Selittämisen vaikeudet

Selittämis- ja perustamisvelvollisuudet ovat tavanomainen tapa vastata tulkinnan vaikeuksiin.[71] Selittämisvaatimuksissa on useita ongelmia. Ensimmäiseksi selittämisvelvollisuudet ovat aina myös teknistä sääntelyä. Ne vaikuttavat siihen, minkälaisia järjestelmiä voidaan ottaa käyttöön, ja ne voivat johtaa siihen, että tarkoituksenmukaisten, tehokkaiden ja luotettavien sovellusten käyttämisestä voi tulla mahdotonta.[72] Hallintolain 58 e §:n asetelma on kuvaava esimerkki siitä, mitä tapahtuu, jos selitettävyydestä pidetään kiinni. Uudistuksessa[73] hallintoviranomaiset velvoitettiin valtiosääntöisistä syistä[74] käyttämään hallintopäätöksenteossa vain selitettäviä sääntöperusteisia päätöksentekojärjestelmiä. Vaatimuksen perustelut ovat ymmärrettäviä[75], mutta samalla ne johtavat siihen, että ainakin jotkin julkishallinnon tekoälytulevaisuudet sulkeutuivat. Hallintopäätöksiä ei vastaisuudessakaan tehdä kuin sääntöperusteisilla automaatiojärjestelmillä, jos sääntely ei muutu.

> Kun suuret kielimallit ovat tulleet markkinoille, voidaan kuvitella, että kielimalli tekisi tulevaisuudessa sekä päätöksen että kirjoittaisi sille perustelut. Tällaisessa asetelmassa selitettävyysvaatimus saattaisi teknisesti täyttyä: päätöksellä on ihmisen ymmärrettävä selitys. Selitettävyyden ongelma siirtyy metatasolle: riittääkö se, että kone, jota emme ymmärrä, kirjoittaa perustelut päätökselle?

Selitettävyysvaatimukset ovat myrkyllisiä myös toisessa suunnassa: ne voivat ohjata tuottamaan selityksiä, joista ei lopulta ole juuri hyötyä. AI-järjestelmissä selityksillä on tavallisesti kaksi tarkoitusperää: järjestelmän oikeuttaminen ja sen kehityksen ohjaaminen. Ensimmäisessä vaihtoehdossa selityksellä pyritään legi-

71. Ks. esim. Corinne Cath, Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376(2133) 2018: 20180080 ja Sandra Wachter – Brent Mittelstadt – Luciano Floridi, Transparent, explainable, and accountable AI for robotics. Science Robotics 2(6) 2017: eaan6080 ja Selbst – Barocas 2018.
72. Ks. esim. Maya Krishnan, Against interpretability: a critical examination of the interpretability problem in machine learning. Philosophy & Technology 33(3) 2020, s. 487–502.
73. Ks. oikeusministeriön Selvityksiä ja ohjeita 14:20: Arviomuistio hallinnon automaattiseen päätöksentekoon liittyvistä yleislainsäädännön sääntelytarpeista ja hallituksen esitys eduskunnalle julkisen hallinnon automaattista päätöksentekoa koskevaksi lainsäädännöksi 145/22 vp.
74. Ks. perustuslakivaliokunnan lausunto hallintovaliokunnalle 7/2019 vp (HE 18/2019 vp laiksi henkilötietojen käsittelystä maahanmuuttohallinnossa ja eräiksi siihen liittyviksi laeiksi).
75. Ks. kotimaisesta keskustelusta Markku Suksi, Förvaltningsbeslut genom automatiserat beslutsfattande – statsförfattnings- och förvaltningsrättsliga frågor i en digitaliserad myndighetsmiljö. JFT 5/2018, s. 329–371; Markku Suksi, Administrative due process when using automated decision-making in public administration: some notes from a Finnish perspective. Artificial Intelligence and Law 29(1) 2021, s. 87–110 ja Ida Koivisto, Thinking inside the box: the promise and boundaries of transparency in automated decision-making. European University Institute 2020 osoitteessa https://cadmus.eui.eu/handle/1814/67272 (vierailtu 1.6.2022).

timoimaan ja oikeuttamaan tekoälyjärjestelmän tuottama tulos. Tavoite asettaa laatuvaatimuksen selitykselle. Selityksen on oltava sellainen, että järjestelmän toiminnan kohteet voivat sen ymmärtää ja he voivat vakuuttua siitä, että järjestelmä toimii niin kuin sen pitääkin, ja lopputulos on oikeutettavissa. Vaikka tulkitsemattomien järjestelmien XAI-tutkimusta[76] on paljon, menetelmien tuotokset eivät aina ole yhteismitallisia oikeuttamistarpeiden kanssa.[77] Esimerkiksi neuroverkoissa niin sanotut post hoc -tulkintamenetelmät voivat tuottaa tietoa siitä, mitä neuroverkon eri kerrokset tai osa-alueet tekevät, mitkä syöte- tai harjoitusdatan ominaisuudet korreloivat tilastollisesti tiettyjen tuotosten kanssa ja mikä konkreettinen vaikutus jollakin syötedatan tilastollisella ominaisuudella on tuotokseen.[78] Tällaiset selitykset eivät tavallisesti ole niitä selityksiä, joita päätösten oikeuttamiseen kaivataan. Jos esimerkiksi hallintopäätös tehtäisiin neuroverkolla, parhaatkaan XAI-käytänteet tai -menetelmät tuskin tuottaisivat oikeudellisissa käytännöissä riittävää perustelua päätökselle. Näissä tapauksissa selitysvaatimuksen hyöty on vähintäänkin kyseenalainen.

Tilanne on usein sama silloin, kun selityksiä halutaan käyttää toiseen selitysten käyttötarkoitukseen eli tekoälykehityksen ohjaamiseen. Tällöin jonkin mekanismin olisi välitettävä selitys normatiivisiksi impulsseiksi kehittäjille. Perinteisessä hallinnollisessa ja oikeudellisessa päätöksenteossa mekanismi on selvä. Selitys on se tekijä, jonka varassa päätöstä tarkastellaan ja sen oikeellisuutta arvioidaan muutoksenhakuprosessissa. Selitys on päätöksen ja koko järjestelmän hyväksyttävyyden mittari. Jos tällaista takaisinkytkentää esimerkiksi riitauttamis-, julkisen tai muun sidosryhmäkeskustelun tai demokraattisen päätöksenteon mekanismien välityksellä[79] ei ole, selitykset kaikuvat kuuroille korville. Selitettävyyden oletetaan tuottavan järjestelmälle legitimiteettiä, vaikka se ei muuta mitään.

Into selitettävyyteen voikin kääntyä fetisismiksi. Selitettävyysvaatimuksista voi tulla lähinnä rituaalisia. Selityksiä on, mutta kukaan ei ymmärrä eikä kuule niitä. On myös tärkeää huomata, että eräissä konteksteissa selitykset eivät auta välittömästi lainkaan. Jos esimerkiksi itsestään ajava auto ajaa jalankulkijan päälle, jalankulkija tuskin hyväksyy ruumiinvammaa, jos hänelle kerrotaan,

76. Ks. esim. Wojciech Samek – Grégoire Montavon – Andrea Vedaldi – Lars Kai Hansen – Klaus-Robert Müller (eds), Explainable AI: interpreting, explaining and visualizing deep learning. Springer 2019.
77. Ks. de Bruijn – Warnier – Janssen 2022; Hacker – Passoth 2022 ja Miriam C. Buiten, Towards intelligent regulation of artificial intelligence. European Journal of Risk Regulation 10(1) 2019, s. 41–59.
78. Post hoc -menetelmistä ks. Lipton 2018 ja Vale – El-Sharif – Ali 2022.
79. Tällaisille takaisinkytkennöille kaavaillaan usein merkittävää roolia tekoälyn hallintastrategioissa. Ks. Inga Ulnicane – Damian Okaibedi Eke – William Knight – George Ogoh – Bernd Carsten Stahl, Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. Interdisciplinary Science Reviews 46(1-2) 2021, s. 71–93.

miksi auto ajoi hänen päälleen.[80] Selityksillä voi toki olla merkitystä, kun järjestelmää kehitetään tai kun onnettomuuden seurauksista käydään oikeutta, mutta selitettävyysvaatimukset voivat olla näistä hyödyistä huolimatta vasta-aiheisia. Jos ei-tulkittavat järjestelmät toimivat selitettäviä paremmin ja niiden luotettavuudesta voidaan varmistua, selitettävyys ei auta mitään, jos se ei auta oikeuttamaan lopputuloksia.

6.3. Teknisen sääntelyn toinen ongelma

Toinen suunta, jossa tulkintavaikeudet vaikuttavat, on tekninen sääntely: sen mahdollisuudet kaventuvat entisestään. Järjestelmien toimintalogiikkaa ei voida muodollisesti eli formaalisti verifioida, kun toimintalogiikkaa ei voida kuvata tutuissa ontologioissa ja tutulla logiikalla.[81] Siksi on ainakin toistaiseksi hyödytöntä vaatia, että formaalin verifioinnin menetelmiä käytettäisiin selittämättömien järjestelmien varmentamiseen. Teknisen sääntelyn kohteeksi jää vain musta laatikko, kone, joka tekee mitä tekee.

7. Syötteet, tuotokset ja epälineaarisuus

7.1. Kun testattavaa on liikaa

Esitin edellä luvussa 5.2., että monimutkaisuuden tuottamista ongelmista voidaan selvitä testaamissääntelyllä. AI-järjestelmien testaamisessa on kuitenkin pulmansa. Jos järjestelmien syöte- tuotosavaruudet ovat laajoja tai järjestelmien toiminta epälineaarista, testaamisesta tulee resurssi-intensiivistä ja teknisesti haastavaa.

Jos järjestelmän syöteavaruus on laaja eli se voi saada laajan kirjon erilaisia syötteitä, testejä tarvitaan paljon. Jos tuotosavaruus on sekin laaja eli järjestelmä voi päätyä lukuisiin lopputiloihin, testaustarve lisääntyy entisestään. Autonomisten ajoneuvojen konenäkö- ja muut sensorijärjestelmät ovat hyviä esimerkkejä. Pelkästään kameroita on tavallisesti useita. Ne tuottavat jatkuvasti miljoonien pikseleiden datavirtoja. Lidar-sensorit tuottavat niin ikään miljoonien pisteiden

80. Vrt. esim. Scott Robbins, A Misdirected Principle with a Catch: Explicability for AI. Minds and Machines 29(4) 2019, s. 495–514.

81. Formaalin verifioinnin menetelmiä kehitetään kuitenkin kuumeisesti, ks. Martin Leucker, Formal Verification of Neural Networks? Formal Methods: Foundations and Applications: 3–72020.

datapilviä. Kun ajoneuvo liikkuu avoimessa maailmassa, erilaisia syötekombinaatioita on käytännössä loputtomasti. Myös mahdollisia tuotostiloja on paljon. Syötteistä pitää tunnistaa potentiaalisesti suuri määrä erilaisia objekteja ja ympäristöjä. Jos kehittäjä haluaa näissä oloissa varmistua siitä, että järjestelmä hahmottaa liikenneympäristön oikein, testejä täytyy tehdä lukemattomia.[82]

Jos järjestelmät ovat epälineaarisia, kuten syväoppimisalgoritmit[83], testaamisen haasteesta tulee kaksin verroin vaikeampi. Lineaarisissa järjestelmissä vasteet syötteiden muutoksiin ovat ennustettavia. Tällöin pistemäinen kulmatapaustestaaminen, jossa määritellään haastavaksi tiedetty tapaus, testataan järjestelmä siinä ja ekstrapoloidaan tulokset lähi- tai vähemmän haastaviin tapauksiin[84], voi olla mielekästä. Tulos tietyillä syötteillä voi antaa luotettavia viitteitä siitä, minkälaisia tuotoksia lähisyötteillä saadaan. Jos järjestelmän käytös on epälineaarista, ekstrapolaatiostrategiat eivät kuitenkaan todennäköisesti toimi. Vaste voi poiketa radikaalisti, vaikka syötteet muuttuisivat vain vähän.

7.2. Katse simulaatioihin

Tekoälyteknologiat ovat kuitenkin läpeensä digitaalisia. Ne elävät tietokoneissa. Tämä tarkoittaa sitä, että suuri osa testaamisesta voi olla virtuaalista. Järjestelmiä ei tarvitse testata todellisessa fyysisessä maailmassa. Skenaariot voidaan pitää tietokoneiden sisällä, jos järjestelmien toimintaa simuloidaan. Simuloinnin mahdollisuus muuttaa sääntelyasetelmaa merkittävästi. Simulaatiotestauksella voidaan tuottaa suljettujen ympäristöjenkin ulkopuolella ainakin jonkinlainen käsitys siitä, miten ja kuinka luotettavasti epälineaariset järjestelmät toimivat.[85]

Simulaatiot näyttävätkin väistämättä tulevaisuuden keskeisimmiltä tavoilta, joilla erilaisten tekoälyjärjestelmien tarkoituksenmukaisesta toiminnasta voidaan

82. Ks. esim. Philip Koopman – Michael Wagner, Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety 4(1) 2016, s. 15–24; Chung Won Lee – Nasif Nayeer – Danson Evan Garcia – Ankur Agrawal – Bingbing Liu, Identifying the operational design domain for an automated driving system through assessed risk. 2020 IEEE Intelligent Vehicles Symposium (IV): 1317–13222020 ja Viljanen 2023.
83. Ks. esim. Lipton 2018.
84. Ks. esim. Daniel Bogdoll – Jasmin Breitenstein – Florian Heidecker – Maarten Bieshaar – Bernhard Sick – Tim Fingscheidt – J. Marius Zollner, Description of corner cases in automated driving: goals and challenges. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW): 1023–10282021 osoitteessa https://ieeexplore.ieee.org/document/9607669/ (vierailtu 30.8.2023); esimerkki tällaisista testausvaatimuksista on UN Regulation No. 157. Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems. ECE/TRANS/WP.29/2020/81.
85. Ks. esim. David Harel – Assaf Marron – Joseph Sifakis, Autonomics: In search of a foundation for next-generation autonomous systems. Proceedings of the National Academy of Sciences of the United States of America 117(30) 2020: 17491–17498.

varmistua. Oletettavasti myös sääntelijät ottavat simulaatiotestausvaatimukset osiksi työkalupakkejaan. Niin näyttää käyvän esimerkiksi ajoneuvosääntelyssä.[86]

Simulointitestauksessa on kuitenkin ongelmansa. Testaaminen on kallista. Testausinfrastruktuurin kehittäminen vaatii määrätietoista työtä ja merkittäviä investointeja. Testiskenaarioiden määrittely vaatii paitsi aikaa ja rahaa myös teknistä ja ennen kaikkea normatiivista asiantuntemusta. Kun skenaarioita määritellään ja hyväksyttävyysmetriikat suunnitellaan, päätetään väistämättä hyvin konkreettisesti siitä, mikä tekoälyn käyttötapauksissa on normatiivisesti merkityksellistä ja mihin hyväksyttävyyden rajat on vedettävä.[87]

Simulaatiotestaaminen voi lisäksi työntää tekoälykehittämisen asetelmia kokonaan uuteen normatiiviseen tilaan. Jos järjestelmien toimintaympäristö pystytään mielekkäästi mallintamaan ja laatimaan kattava skenaariokatalogi, järjestelmien tuottamia lopputuloksia pystytään tarkastelemaan laadullisesti uudella tavalla. Esimerkiksi A/B-simulaatioissa eri järjestelmäversioita ja niiden tuottamia lopputuloksia voidaan verrata keskenään. Kun simulaatiot tuottavat tietoa eri vaihtoehtojen todennäköisistä lopputuloksista, kehittäjät pääsevät ja joutuvat tekemään suunnitteluvalintoja tietoisina niiden todennäköisistä vaikutuksista. Tämä voi johtaa vastuusääntelyn kannalta ongelmallisiin tilanteisiin: esimerkiksi autonomisen ajoneuvon kehittäjät voivat tietää, kuinka monta onnettomuutta ajoneuvon eri versiot todennäköisesti aiheuttavat ja millaiset tienkäyttäjäryhmät ovat haavoittuvassa asemassa eri versioissa. Valintojen moraaliset ulottuvuudet tulevat käsinkosketeltaviksi ja hyvin välittömiksi. Pystytäänkö esimerkiksi rikosoikeudessa vastaamaan tällaisen uudella tavalla tilastollisesti tietoisen riskinoton haasteeseen?[88] Jää nähtäväksi.

8. Indeterministisen koneen ongelmat

Simulaatioiden rajat tekoälyjärjestelmien hallintakeinona tulevat vastaan, jos järjestelmät eivät toimi kuin deterministiset koneet, joka kerta samalla ennakoitavalla tavalla.

Indeterminanssia on kahta laatua, joista toinen ei ole täysipainoista, vahvaa indeterminanssia. Heikossa indeterminanssissa tekoälyjärjestelmien vasteet vaihtelevat samoilla syötteillä mutta vasteiden jakauma on stabiili. Tällainen jär-

86. World Forum for Harmonization of Vehicle Regulations, New assessment/test method for automated driving (NATM) – master document 2021 osoitteessa https://unece.org/sites/default/files/2021-04/ECE-TRANS-WP29-2021-61e.pdf (vierailtu 31.5.2022).

87. Ks. esim. Viljanen 2023.

88. Teemasta on kirjoitettu vähän. Ks. kuitenkin esim. Kenneth W. Simons, Statistical Knowledge Deconstructed. Boston University Law Review 92(1) 2012, s. 1–88 ja Luoto 2022.

jestelmä ei tuota aina samoilla syötteillä samaa tuotosta mutta toimii silti johdonmukaisesti, koska eri vasteiden todennäköisyys voidaan mitata tai arvioida. Jos vasteiden jakauma on ennustettavissa ja vakaa, järjestelmän toimintaa voidaan simuloida mielekkäällä tavalla. Simulointi on kuitenkin entistäkin vaikeampaa, koska simuloinnissa on syöteavaruuden lisäksi katettava myös vasteiden kirjo riittävällä tavalla.

Vahvassa indeterminanssissa vasteilla ei ole stabiilia tai ennustettavaa jakaumaa. Vastetta, joka syötteestä seuraa, ei voida ennustaa mielekkäällä tavalla. Vahvan indeterministisissä tekoälyjärjestelmissä simulointivarmennus jää väistämättä puutteelliseksi. Vaikka koko syöteavaruus voitaisiin kattaa, järjestelmän toimintaa ei voida aukottomasti simuloida. Tällöin joudutaan sietämään sitä, että järjestelmissä osa vasteista jää kartoittamatta.

> Ei ole selvää, kuinka yleisiä indeterministiset järjestelmät tällä hetkellä ovat. Joissain järjestelmissä saattaa olla rakenteita, jotka tuottavat järjestelmien tuotoksiin satunnaisuutta. Esimerkiksi erilaiset generatiiviset kielimallit vaikuttaisivat tuottavan osin sattumanvaraisia tuloksia.

Jos tekoälyjärjestelmä on monimutkainen, sen tulkitseminen ei onnistu, ja jos se on vielä toiminnaltaan indeterministinen, sääntelymahdollisuudet kuivuvat pitkälti kasaan. On hyvin vaikeaa, jos ei käytännössä mahdotonta, varmistua ennalta siitä, että järjestelmä toimii tarkoituksenmukaisella tai hyväksyttävällä tavalla. Sääntelyvaihtoehdoiksi jäävät käytännössä vain kiellot. Jos riskit hahmotetaan liian suuriksi, indeterministiset tekoälyjärjestelmät on kiellettävä.

9. Dynaamisen koneen haaste

9.1. Nykyinen tekoäly ei ole dynaamista

Yksi yleisimmistä tekoäly-ymmärrysharhoista on se, että koneoppivat järjestelmät ”oppivat” jatkuvasti eli muuttuvat käytön aikana. Syy on kielenkäytössä. Tekoälytutkimuksessa ja tietojärjestelmätieteessä puhutaan oppimisesta ja autonomiasta, mutta sanat eivät tarkoita sitä kuin juristi äkkiseltään ymmärtäisi. Tilanne on vaarallinen: juristeilla insinöörien ja matematiikkojen sanat kääntyvät usein ajatukseksi, että järjestelmäinstanssi oppisi jatkuvasti kokemastaan ja havaitsemastaan. Tällöin tekoälyjärjestelmät hahmottuvat kokonaan arvaamattomiksi ja oikukkaiksi. Ne voivat tehdä mitä vain.[89]

89. Ks. esim. Karnow 2016, s. 52; Chopra – White 2011, s. 10 ja Barfield 2018.

Tosiasiassa jatkuvasti oppivia järjestelmiä ei kuitenkaan juuri ole olemassa: lähestulkoon mitkään tekoälyjärjestelmät eivät opi mitään käytön aikana, eikä niissä ole teknisiä rakenteita, jotka muuttaisivat tekoälyjärjestelmän algoritmeja käytön aikana. Esimerkiksi autonomisen ajoneuvon ohjausjärjestelmän koneoppimiskomponentit[90] koulutetaan ”laboratorioissa” keskitetysti ja järjestelmätasolla.[91] Kehittäjät keräävät dataa, esikäsittelevät tai esikäsittelyttävät sen ja sitten käyttävät erilaisia kone- ja syväoppimismenetelmiä järjestelmän algoritmien kehittämiseen. Tämän ”kouluttamisen” jälkeen algoritmit ja niistä koostuvat järjestelmät validoidaan ja testataan. Autojen toimintaa simuloidaan ja niitä ajetaan koeradoilla ja yleisessä liikenteessä, jotta voidaan varmistua siitä, että järjestelmät toimivat riittävän hyvin. Vasta validoinnin ja testaamisen jälkeen järjestelmä siirretään ”tuotantokäyttöön” yksittäisiin instansseihin vakaina ohjelmistopäivityksinä.

Useimmat järjestelmät eivät siis ole käytön aikana dynaamisia saati vahvasti autonomisia. Päinvastoin: ne ovat vakaita ohjelmistoja, joiden toimintaperiaatteet ja algoritmit eivät muutu omia aikojaan. Siksi tietojenkäsittelytieteen kielellä oppivat järjestelmät toimivat kuin mikä tahansa ohjelmisto ja kuin koneet: ne reagoivat ympäristösyötteisiin koodinsa määräämällä tavalla.

Päästään tärkeään seikkaan. On vaarallista rakentaa sääntely sen oletuksen varaan, että koneet ovat dynaamisia ja oppivat ja muuttuvat jatkuvasti. Tämä kehystys johtaa keskustelun tieteiskirjallisuuden tilaan. Ja mikä tärkeintä, se luonnollistaa mielipuolisen vaarallisen teknologisen tulevaisuuden ja tekee siitä ikään kuin vääjäämättömän. Sen sijaan, että puhumme dynaamisista tekoälyjärjestelmistä korkean riskin käyttötapauksissa ja pohdimme, miten niiden kanssa voi elää, olisi syytä pitää huoli siitä, että tällaisia tulevaisuuksia ei pääse syntymään.

9.2. Elo arvaamattoman koneen kanssa

Voi tietysti olla, että päädymme tulevaisuudessa vastakkain dynaamisten oppivien koneiden kanssa. Tutkijat ovat jo nyt kehittäneet menetelmiä, joilla koneet voitaisiin saada oppimaan koko elinkaarensa uudesta datasta, jota ne keräävät käytön aikana. Tällaisessa elinikäisessä oppimisessa (lifelong learning)[92] järjes-

90. Ks. esim. Sorin Grigorescu – Bogdan Trasnea – Tiberiu Cocias – Gigel Macesanu, A survey of deep learning techniques for autonomous driving. Journal of Field Robotics 37(3) 2020, s. 362–386.
91. Ks. esim. Jianyu Chen – Jingliang Duan – Yang Guan – Qi Sun – Yuming Yin – Shengbo Eben Li, Self-learning decision and control for highly automated vehicles, s. 307–330 teoksessa Yi Lu Murphey – Ilya Kolmanovsky – Paul Watta (eds), AI-enabled technologies for autonomous and connected vehicles. Springer 2023.
92. Ks. Zhiyuan Chen – Bing Liu, Lifelong machine learning. Second edition. Morgan & Claypool 2018.

telmä käyttää ennalta määriteltyjä koneoppimismenetelmiä datamassoihin, joita sen toiminnassa kertyy. On myös teoriassa mahdollista, että instanssi voisi ”oppia oppimaan” eli valita valmiista menetelmistä sopivan tai kehittää itsekseen uusia koneoppimismenetelmiä. Menetelmiä tutkitaan niin sanotun automatisoidun koneoppimisen (AutoML) tutkimuksessa.[93] Oma kysymyksensä on, miten mielekästä käytönaikainen elinikäinen oppiminen saati automaattinen koneoppiminen käytännössä olisi. Nykyiset koneoppimismenetelmät ovat sekä laskenta- että energiaintensiivisiä. On todennäköistä, että instanssitason oppiminen ei olisi edes taloudellisesti kannattavaa, saati turvallista tai muutoinkaan mielekästä.

Jos koneista tulee dynaamisia, päädytään oikeudellisen hallinnan viimeiselle rajalle. Vaihtoehdot ovat vähissä. Ensimmäinen vaihtoehto on yrittää teknisillä keinoilla varmistaa, että kone ei opi älyttömiä. Testaussääntely vaikuttaa ainoalta mielekkäältä tavalta pyrkiä varmistamaan oppivan järjestelmän asianmukainen oppiminen. Sääntelyllä voitaisiin edellyttää, että koneihin rakennetaan toiminnallisuuksia, jotka varmentavat, että oppimistulokset ovat asianmukaisia ennen kuin ne siirretään tuotantokäyttöön. Tällainen automaattinen verifiointi ja validointi on ainakin käsitteellisesti mahdollista. Kuinka käytännöllistä ja helppoa järjestelyt on toteuttaa, jää nähtäväksi.

Toinen vaihtoehto vie metatasolle ja tieteiskirjallisuuden maailmaan. Voimme ainakin kuvitella, että oppivaan koneeseen voitaisiin rakentaa toiminnallisuuksia, jotka pitävät huolen siitä, että koneen toimintaperiaatteet tai sen tuottamat lopputulokset ovat eettisesti hyväksyttäviä ja laillisia. Miten toiminnallisuudet toteutettaisiin, jää epäselväksi.

Oppivan koneen sääntely on siis haasteellista ja näyttää mahdolliselta vain, jos käytettävissä on teknisesti hyvin kehittyneitä varmennusmenetelmiä tai tapoja motivoida koneita hyväksyttäviin tarkoitusperiin. Jos tällaisia menetelmiä ei ole, riski siitä, että järjestelmien oppimistulokset ovat epätoivottavia, on aina olemassa. Tästä syystä oppivia järjestelmiä ei tulisi käyttää ainakaan turvallisuuskriittisissä konteksteissa.

10. Menikö juna jo?

Tekoälysääntely on lähdössä hiljalleen käyntiin. EU:n tekoälysäädöksessä päästäneen maaliin ennen kevään 2024 EU-parlamenttivaaleja, samoin ehkä tekoälydirektiivissä. Sääntelytulevaisuus ei kuitenkaan näytä kovin valoisalta. Tekoälyn

93. Ks. esim. Rafael Barbudo – Sebastián Ventura – José Raúl Romero, Eight years of AutoML: categorisation, review and trends. Knowledge and Information Systems 2023 osoitteessa https://link.springer.com/10.1007/s10115-023-01935-1 (vierailtu 3.9.2023).

haaste oikeusjärjestelmälle on vakava: sääntelyssä vähätöiset ja tehokkaat keinot ovat vähissä.

Vastuusääntelyn yleiset opit nitisevät liitoksissaan kahdessa toistensa kanssa ristiriitaisessa suunnassa. Ensimmäisessä suunnassa vastuusääntelyn perinteisten kulmakivien ongelmanratkaisuvoima näyttää olevan katoamassa. Ihmiset ymmärtävät prosesseja, jotka panevat liikkeelle entistä huonommin ja yksittäisiin päätöksiin tiivistyy entistä mittavampia tulevaisuusketjuja. Ajatellaan esimerkiksi autonomista ajoneuvoa. Ohjausjärjestelmäversion hyväksyjä ei välttämättä ymmärrä, mitä on laittamassa liikkeelle, mutta valinnan vaikutukset skaalautuvat lähes välittömästi miljooniin instansseihin eikä prosessia voi ehkä edes vakuuttaa.[94] Toisessa suunnassa tilanne on päinvastainen. Kehitys- ja validointikäytännöt voivat tuottaa ennennäkemättömän määrän kvantitatiivista simulointitietoa järjestelmien toiminnan todennäköisistä seurauksista. Jos ja kun tietoa on paljon, valinta ei tapahdukaan enää epävarmuuden olosuhteissa vaan eräänlaisessa tilastollisen varmuuden tilassa. Ainakaan nykyinen vastuusäännöstö ei kykene vastaamaan tällaisiin haasteisiin.

Teknisessä sääntelyssä joudutaan niin ikään mukautumaan uuteen todellisuuteen. Perinteinen teknologinen käskysääntely tuskin toimii tekoälyteknologioiden tapauksessa, kun vaihtoehtoisia tapoja saavuttaa sama toiminnallisuus on lähes loputtomiin. Kun teknologioiden toimintatavat ovat vielä enenevässä määrin selittämättömiä ja järjestelmät käsittämättömän monimutkaisia, moni sääntelypolku sulkeutuu. Selittämisvelvollisuuksilla ei ole enää aina tilaa. Formaalit verifiointivaatimukset käyvät turhiksi, koska järjestelmiä ei voida verifioida. Sääntelyssä joudutaankin kääntymään joko kohti tulosperusteista metasääntelyä tai sitten kieltämään tekoälyn käyttäminen tietyissä käyttöympäristöissä kokonaan. Testaaminen simuloimalla on tulevaisuuden sääntelyä. Sen ongelmia ei sovi vähätellä. Kun tieto todennäköisestä tulevaisuudesta ja valintojen hyödyistä ja kustannuksista tarkentuu, sääntelyvaatimusten asettamisesta näyttäisi väistämättä tulevan normatiivisesti tulenarkaa. Joudumme päättämään, kuka kärsii ja miten.

Jossakin horisontissa häämöttää tekoälyteknologioiden tieteisfiktiotulevaisuus, jossa teknologiat voivat irrota oikeudellisesta hallinnasta. Vähäisessä mitassa merkkejä on näkyvissä. Jos tekoälykehittäjät luovat dynaamisia järjestelmiä ja käyttävät niitä korkean riskin käyttötapauksissa, ollaan hallinnan viimeisellä rajalla. Silloin ainoa toivo on, että tekoälyn hallinnan menetelmät voidaan automatisoida ja riittävät varokeinot upottaa järjestelmiin.

On kuitenkin tärkeää huomata, että sääntelytavoilla, joita tarkastelin edellä, voidaan pääasiassa hallita vain tekoälyteknologioiden välittömiä seurauksia. Välillisten vaikutusten hallintaan välineitä ei juuri ole, vaikka tekoälyteknо-

94. Ks. esim. Mika Viljanen, Robotteja vakuuttamassa: autonomiset alukset esimerkkinä. Lakimies 7–8/2018, s. 954–974.

logiat näyttävät johtavan väistämättä moniin sosiaalisiin transformaatioihin. Työpaikkoja katoaa, demokratia joutuu uhatuksi, ihmisyys muuttuu fundamentaalisesti. Näiden haasteiden edessä sääntelijät näyttävät antautuvan taistelutta: poliittista tahtoa välillisten haittojen hallintaan ei ole. Pikemminkin päinvastoin: esimerkiksi EU:n tekoälysäädös on viritetty erittäin innovaatioystävälliseksi sääntelyinstrumentiksi. Säädöksen 5 artiklassa ei kiellettäisi kuin kourallinen rankimpia tekoälyn väärinkäyttötapauksia. Muut säädöksessä asetettavat sääntelyvaatimukset ovat hyvin kevyitä, pääasiassa menettelyllisiä ja siinä määrin epämääräisiä, että niiden teho jää kyseenalaiseksi. Sääntelyvaatimukset kohdistuvat lisäksi vain suhteellisen pieneen joukkoon niin sanottuja korkean riskien tekoälyjärjestelmiä. Muut matalan riskin järjestelmät jätetään vapaaehtoisten käytännesääntöjen varaan ja jäsenvaltioita kielletään asettamasta järjestelmille kansallisesti lisäsääntelyvaatimuksia.

Jotta kuva ei olisi toivoton, on syytä pitää mielessä, että tekoälyteknologioissa on uskomaton potentiaali hyvään. Saatamme myös pelätä koneita turhaan ja siinä sivussa sietää ihmisiltä käyttäytymistä, vahinkoja ja haittoja, jotka tekoälyteknologiat voisivat vaivatta poistaa.[95] Juna on silti hyvää vauhtia lähdössä asemalta. Niin on tietysti ollut koko modernin ajan. Kehitys on kehittynyt, ja yhteiskunnat ovat joutuneet sopeutumaan. Nyt muutosvauhti vaikuttaa kuitenkin ainutlaatuisen nopealta.

95. Ks. ns. robofobiasta Andrew Keane Woods, Robophobia. University of Colorado Law Review 93(1) 2022, s. 51–114.

Has the train already left the station? Options for regulating artificial intelligence

MIKA VILJANEN, LL.D., Professor, University of Turku

The article examines the regulatory options for AI technologies. It is argued that AI regulation proposals and literature frame AI in terms that cause misunderstanding in the discussion/debate and suggest it is about unpredictable and autonomous learning systems that are difficult if not impossible to control. However, this is not the case. The article proposes six new articulations for regulation-relevant AI properties. The properties are technical agency, complexity, interpretability, non-linearity, and the extent of input and output spaces, (in) determinacy and dynamism of the systems. When these articulations guide analysis, new regulatory problems and opportunities arise. The article then moves on to explore the regulatory tools that could be used to address the immediate adverse effects of each of these properties. The outlook is a cause for optimism but also concern. The immediate adverse effects of AI technologies can be addressed, but the solutions are not easy. Meta or process regulation, simulation-based performance regulation and appropriate explanation and transparency regulation may help in controlling the potential harm of AI. Simultaneously, care must be taken to ensure that indeterministic and dynamic systems are not allowed to enter safety-critical environments. Regarding the control of indirect damage, there are few options. The means for controlling the potential harm of AI are few in number and there is little political will for regulatory projects in sight.