ROBOTIT JA TEKOÄLY MORAALISINA HUOLENAIHEINA, TOIMIJOINA JA NEUVONANTAJINA

ARTO LAITINEN

Voivatko robotit tai tekoälyjärjestelmät tulevaisuudessa olla moraalisia toimijoita tai moraalisten toimijoiden neuvonantajia? Jos niin on, millaista moraalikoodistoa niiden tulisi noudattaa? Monet huolenaiheet liittyen robotiikkaan tai tekoälyihin ovat täysin riippumattomia näistä kysymyksistä: järjestelmät voivat saada aikaan vahinkoa, vaikka eivät olekaan moraalisia toimijoita.

Robotit ja tekoäly moraalisina huolenaiheina

Monet huolenaiheet liittyen robotiikkaan tai tekoälyihin ovat täysin riippumattomia siitä, ovatko ne moraalisia toimijoita. Systeemi voi saada aikaan vahinkoa, vaikka se ei olekaan moraalinen toimija. Sen sijaan vain moraalisen arvioinnin piiriin kuuluvat toimijat (kuten tyypilliset aikuiset ihmiset) toimivat moraalisesti *väärin*, jos esimerkiksi aiheuttavat turhaa kärsimystä eläimille. Jos saman kärsimyksen aiheuttaa kaatuva puu, niin se ei toimi moraalisesti väärin: puuta ei voi pitää moraalisesti vastuullisena tahona.

Kun pieni lapsi sairastuu ja kuolee, saatetaan sanoa, että se on ”epäoikeudenmukaista”, ”julmaa” tai ”vääryys”, vaikka kirjaimellisesti mikään toimija ei olisikaan toiminut väärin, julmasti tai epäoikeudenmukaisesti. Tavallisesti tässä puhetavassa ei ole mitään harhaanjohtavaa, mutta kun tarkastellaan robotteja tai tekoälyä hyödyntäviä järjestelmiä, ollaan lähempänä väärinkäsityksiä.

Tekoälyn, robottien ja algoritmien vaarat koskevat keskeisesti tällaisia epätoivottavia seurauksia. Esimerkiksi Cathy O’Neil tuo esiin erilaisia yhteiskunnallisia haittoja algoritmien käytössä: töihinotossa, työntekijöiden arvioinnissa, poliisien ja sairaaloiden toiminnassa jne. Lisäksi erilaisten automatisoitujen asejärjestelmien, itseään ohjaavien ajoneuvojen tai mahdollisten seksirobottien toivottavia tai epätoivottavia seurauksia voidaan arvioida (ks. myös Laakasuon ja Palomäen teksti edellä). Moraalista vastuuta on ihmisillä, jotka näitä suunnittelevat ja käyttävät. Tällä hetkellä painavimmat huolet eivät siis lainkaan riipu siitä, ovatko tai voivatko robotit tai tekoälyjärjestelmät tulevaisuudessa olla moraalisen toimijuuden ehdot täyttäviä.

Voivatko robotit tai tekoäly tulevaisuudessa olla moraalisia toimijoita?

Hieman yksinkertaistaen voidaan erottaa kaksi näkökulmaa robotteihin: yhtäältä toimijoiden kokemuksellinen ja osallistuva arkinäkökulma sekä toisaalta selityksiin tähtäävä tieteellinen näkökulma. Näistä näkökulmista keinotekoisten kognitiivisten systeemien ja ihmisten ero näyttäytyy hyvin erilaisena. Hans Jonas (1966) kutsuu näitä näkökulmia fenomenologiseksi ja kyberneettiseksi. Näkökulmien kiistely sinänsä on hedelmätöntä, ne tulisi molemmat ottaa huomioon. Wilfrid Sellars (1962) näkeekin filosofian tehtäväksi luoda stereoskooppista kuvaa, jossa arkinen ”ilmikuva” ja tieteellinen maailmankuva tuodaan yhteen. On kuitenkin tärkeä erottaa, kumman näkökulman edellyttävillä käsitteillä operoidaan.

Arkinäkökulmasta ihmisen ja robotin välinen ero vaikuttaa selvältä. Ihminen on elävä, tunteva, tietoinen, itsetietoinen, harkintaan ja perusteiden punnintaan kykenevä olento, jolla on kyky välittää ja tuntea moraalisia tunteita. Monimutkaisillakaan koneilla näitä piirteitä ei ole, vaikka ne olisivat itseohjautuvia. Ne voivat vain *simuloida* näitä piirteitä

ulkoisessa käyttäytymisessään (Seibt 2017). Tulevaisuuden visio roboteista, jotka olisivat moraalisia toimijoita, edellyttää tästä näkökulmasta laadullista hyppyä, ja on epäselvää, voivatko koneet tulevaisuudessakaan saada näitä piirteitä – esimerkiksi tuntea kipua tai moraalisia tuntemuksia.

Toisesta näkökulmasta laadullista hyppyä ei tarvita. Ihmiset ja koneet ovat jo nyt, vaikkakin yksityiskohdissaan hyvin erilaisia, periaatteessa samanlaisia systeemejä. Jossain sopivassa mielessä koneet ovat ”tietoisia”, ”toimijoita”, ”pyrkiviä”, ”kehollisia”, ”päätöksentekijöitä”, ”autonomisia” – esimerkiksi itseään ohjaavat kulkuneuvot jossain minimaalisessa mielessä ovat näitä kaikkia. Kokemuksellinen tietoisuus ei ole merkitsevä seikka näiden tietojenkäsittelysysteemien toiminnan selittämisessä. Jos laadullista eroa ei ole, voidaan ajatella ihmisten ja tekoälyn eroa esimerkiksi määrällisen laskentatehon kasvun näkökulmasta. Myöskään moraalista toimijuutta ei tästä näkökulmasta pidä ajatella jonain laadullisesti erityisenä seikkana. Monien mielestä tällöin pikemminkin hukataan aito moraalinen toimijuus – tarvitaan koettua fenomenologista näkökulmaa, jotta edes tavoitetaan kyseinen ilmiö.

Tekoäly moraalisena neuvonantajana?

Tekoäly voi jo nyt toimia oikeusapulaisena, käydä läpi oikeustapauksia, niissä esitettyjä argumentteja ja päätöksiä: auttaa ihmisiä päättämään, mitä tehdä. Voisiko sama toimia moraalissa: moraalisen dilemman kohdatessa ”kilauta keinoälylle” tai tee haku kuvitteellisella ”Google Morals”-ohjelmalla (Howell 2012)?

Tällaisen kehittelyssä ensimmäisen vaiheen voisivat muodostaa esimerkiksi tutkimusetiikan tai sairaaloiden eettiset toimikunnat, jotka tuottavat dokumentteja eettisistä kannanotoista. Päämääränä voisi olla laajempikin arkielämän tietokanta. Tietokannan haasteellisuus tulee ilmi, kun tarkastellaan neljän tasoista moraalista tietoa.

Yleisimmät moraaliteoriat (seurausetiikka, kantilainen velvollisuusetiikka, kontraktualismi eli sopimusteoreettinen etiikka, ehkä hyve-etiikka) pyrkivät muotoilemaan yleisen, kaikkiin tilanteisiin soveltuvan poikkeuksettoman periaatteen: ”toimi aina niin, että…”. Mikä periaate olisi oikea? Asiaa on tarkasteltu pitkään ja hartaasti, eikä asiantuntijoiden kesken ole luvassa konsensusta. Tästä ei tekoälyn ohjelmoijan kannattane aloittaa, vaan antaa tällaisen poikkeuksettoman periaatteen syntyä, jos on syntyäkseen. Toisaalta voi ajatella, että ohjelma antaisi useampia vastauksia: utilitarismin mukaan teko on väärin, kantilaisen velvollisuusetiikan mukaan oikein jne.

Huomattavasti kiistattomampia ovat keskitason *prima facie* -velvollisuudet koskien yksittäisiä tekoluokkia. Pluralistisia keskitason velvollisuuksien listoja ovat esimerkiksi Raamatun 10 käskyä, W. D. Rossin (1930) lista tai bioetiikan raamatun, Tom Beauchampin ja James Childressin (2013) neljä periaatetta (autonomia, vahingoittamisen kielto, hyvinvoinnin edistäminen, oikeudenmukaisuus). Näistä kannattanee tietokannan rakentaminen aloittaa (ks. Anderson ja Leigh Anderson 2011). Yksinään tarkasteltuina, ottamatta konflikteja huomioon, nämä periaatteet eivät useinkaan ole kiistanalaisia: oikeudenmukaisuus on hyvä asia, epäoikeudenmukaisuus on paha asia. (Kiistat koskevat usein metafyysisiä taustaoletuksia: ”kunnioita Jumalaa” ei ole mielekäs käsky, ellei Jumalaa ole olemassa.)

Kolmanneksi voidaan konfliktitapauksissa kaivata keskitason periaatteiden keskinäistä painoarvoa koskevia ”prioriteettiperiaatteita”. Näitä ovat esimerkiksi F. M. Kammin (2016) periaate, jonka mukaan yhden henkilön saa uhrata viiden vuoksi vain, jos uhria ei käytetä kausaalisesti muiden pelastamisen välineenä. Michael ja Susan Anderson (2011) tarkastelevat, saako potilas kieltäytyä lääkityksestä. He argumentoivat, että autonomia on tärkein periaate, sen jälkeen vahingoittamisen kielto ja sitten hyvinvoinnin edistämisen vaatimus. Nämä prioriteettiperiaatteet saattavat kuitenkin olla periaatteessakin saavuttamattomia, jos ajatellaan, että ”ratkaisu syntyy tilannekohtaisessa havainnossa” (Ross 1930). Joskus valehtelu toiselle on sallittua toisen hyvinvoinnin vuoksi, joskus ei; tämän näkeminen vaatii tilannekohtaista harkintaa. Toisaalta, periaatteessa tekoäly voisi havaita toistuvia kaavoja, implisiittisiä prioriteettiperiaatteita, jos sille syötettäisiin riittävä määrä tapauskohtaisia arvostelmia.

Neljännen analyysitason muodostavatkin tilannekohtaiset arvostelmat, kun *kaikki* relevantti tilanteessa on otettu huomioon. Jokainen toimintavaihtoehto jokaisessa tilanteessa on kaiken kaikkiaan moraalisesti sallittu (oikein) tai ei (väärin).

Näistä muodostuu valtava moraalinen kartta: jokaisen tilanteen jokaisen toimijan jokainen toimintavaihtoehto tulisi olla mukana. Yhden lisäpiirteen lisääminen voi keikauttaa tilanteen moraalisesti päälaelleen (jos napista painamalla laukaistaan lisäksi ydinpommi), joten arvostelmilla, joissa kaikki oleellinen ei ole mukana, ei tee mitään. Saattaa hyvin olla, että luotettavaa tietokantaa koskien tämän tason arvostelmia ei voi luoda. Paras apu, mitä tekoälyltä voi saada tilannekohtaisissa arvostelmissa, saattaa olla mahdollisesti relevanttien piirteiden tarkastuslista, kuten ”Kohdellaanko kaikkia yhdenvertaisesti?”, ”Onko kyse jonkun luottamuksen pettämisestä?” ja ”Onko otettu huomioon ympäristövaikutukset?”.

Onko aihetta optimismiin?

Optimismi koskien sitä, saadaanko tekoälystä moraalista neuvonantajaa, riippuu osittain siitä, miten nähdään eritasoisten periaatteiden merkitys moraalissa: antiteoreettiset lähestymistavat pitävät tehtävää periaatteessakin mahdottomana, monet muut vain käytännössä toivottomana. Optimismi koskien sitä, saadaanko roboteista moraalisia toimijoita, riippuu siitä mitä tarkoitetaan – painotetaanko arkifenomenologista vai tieteellistä näkökulmaa. Pessimismi koskien robotteja ja tekoälyä huolenaiheina sen sijaan ei riipu näistä: järjestelmät voivat saada aikaan vahinkoa, vaikka eivät olekaan moraalisia toimijoita.

Kirjallisuus

Anderson, Michael ja Leigh Anderson, Susan (toim., 2011) *Machine Ethics*, Cambridge UP.

Beauchamp, Tom ja Childress, James (2013), *Principles of Biomedical Ethics*, 7. p., Oxford: Oxford UP.

Howell, Robert (2014), ”Google Morals, Virtue, and the Asymmetry of Deference”, *Nous* 48(3).

Kamm, F. M. (2016) *The Trolley Problem Mysteries,* Oxford: Oxford University Press.

O’Neil, Cathy (2017) *Matikkatuhoaseet.* Terra Cognita.

Ross, W. D. (1930) *The Right and The Good.* Oxford: Oxford University Press.

Seibt, Johanna (2017) ”Towards an Ontology of Simulated Social Interaction: Varieties of the ”As If” for Robots and Humans”. Teoksessa *Sociality and Normativity for Robots,* toim. J. Seibt ja R. Hakli. Springer.

Sellars, W. (1962) ”Philosophy and the Scientific Image of Man”, teoksessa *Frontiers of Science and Philosophy*, toim. Robert Colodny. Pittsburgh: University of Pittsburgh Press, 35–78.

Kirjoittaja on filosofian professori Tampereen yliopistossa ja on mukana Suomen Akatemian Strategisen tutkimuksen neuvoston rahoittamassa hankkeessa ”Robotit ja hyvinvointipalveluiden tulevaisuus” (ROSE).