•

Voiko robotti olla moraalisesti vastuullinen toimija?

AKU VISALA

TIIVISTELMÄ Ihmistä muistuttavien ja ihmisen käyttäytymistä simuloivien robottien moraalinen status ei ole selvä. Artikkelissa kiinnitetään huomioita erityisesti niihin käytäntöihin, joista moraalinen vastuullisuus koostuu. Näitä ovat esimerkiksi rankaiseminen, moittiminen, kehuminen ja palkitseminen. Missä määrin moraalisen vastuullisuuden asenteet ja käytännöt ovat oikeutettuja robotteja kohtaan? Vastuupessimistien mukaan robotteja ei tulisi koskaan pitää moraalisessa vastuussa, koska näiltä puuttuu jokin kyky, joka on välttämätön moraaliselle toimijuudelle – esimerkiksi tietoisuus, autonomia tai vapaa tahto. Optimistit puolestaan ajattelevat, että robotit voisivat ainakin periaatteessa kuulua vastuuasenteidemme piiriin. Artikkeli esittelee ensin keskustelua moraalisesta vastuusta ja sen perusteluista. Tämän jälkeen se tarkastelee kahta vastuupessimistien argumenttia, joista ensimmäinen koskee tietoisuutta ja toinen autonomiaa. Esitettyään kriittisiä huomioita näistä argumenteista se hahmottelee varovaisen optimistista kantaa, jonka mukaan robotit eivät ehkä kykene täysimääräiseen vastuullisuuteen, mutta ne kuitenkin voisivat olla sopivia kohteita ainakin joillekin vastuuasenteille.

ASIASANAT Moraalinen vastuullisuus, robotti, tahdonvapaus, tietoisuus, autonomia.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Euroopan unioni ja monet sen jäsenmaat ovat julkaisseet omat tekoälyä koskevat strategiansa.[1] Näissä strategioissa korostetaan sitä, kuinka ihmiset ovat itse vastuussa teknologian kehityksestä ja sen vaikutuksesta yhteiskuntaan. Teknologista kehitystä ei tule jättää sattuman varaan, vaan ihmisten tulee ottaa siitä vastuu. Strategioiden mukaan ei saa syntyä tilanteita, joissa vastuu häviää tai hajoaa tavalla, joka tekee sen kohdistamisen mahdottomaksi. Strategioiden huolena on tilanne, jota voitaisiin kutsua ”vastuuvajeeksi” (engl. responsibility gap). Vastuuvaje on tilanne, jossa teknologia aiheuttaa merkittävää haittaa mutta samaan aikaan on epäselvää, kuka tästä on oikein vastuussa.[2] Älykkäällä teknologialla on potentiaali synnyttää aiempia suurempia vastuuvajeita, koska se on yhä enemmän autonomista. Autonomialla tarkoitetaan tässä yhteydessä sitä, että jokin teknologia toimii yhä itsenäisemmin tekijästään riippumatta ja se kykenee oppimaan niin uusia keinoja kuin päämääriäkin.

Vastuuvajeen tekee vielä hankalammaksi se, että älykäs teknologia on osa suurta toimijoiden verkostoa. Kun itseohjaava auto joutuu onnettomuuteen, onko vastuussa esimerkiksi auton suunnittelija, ohjelmiston koodaaja, ohjelmiston päivittäjä, auton valmistaja, autokauppias, käyttäjä vai mekaanikko? Tulisiko meidän laskea syyllisten joukkoon mukaan myös älykäs teknologia itse, tässä tapauksessa itseohjaava auto? Mitä itseohjaavan auton moittiminen, paheksuminen tai rankaiseminen voisi edes tarkoittaa? Jos järjestelmä voisi olla itse vastuussa toiminnastaan, niin valmistajat, käyttäjät kuin ohjelmoijatkin voisivat pahimmassa tapauksessa välttää vastuun järjestelmän toiminnan aiheuttamista seurauksista.

Tarkastelen tässä artikkelissa filosofista keskustelua siitä, missä määrin tekoälyä käyttävät teknologiat voisivat olla moraalisesti vastuullisia omista teoistaan. Keskityn inhimillistä toimintaa matkiviin robotteihin, koska kysymys vastuullisuudesta herää vain silloin, kun kone voi tulla osaksi inhimillisen sosiaalisen ja moraalisen elämän verkostoa. Robotilla tarkoitan mahdollista autonomista oliota, joka kykenee käsittelemään tietoa eri elämänalueilta (ongelmanratkaisu, sosiaalinen informaatio) ja joka voi toimia fysikaalisessa ja sosiaalisessa maailmassa.[3] Robottietiikan piirissä on viime vuosien aikana syntynyt vilkas keskustelu robottien mahdollisesta moraalisesta toimijuudesta sekä moraalisesta vastuullisuudesta.[4]

Sivuutan kysymykset teknologisen kehityksemme tasosta ja siitä, mihin suuntaan tekoäly ja robotiikka saattavat kehittyä lähivuosina. Lähestymistapani on filosofinen, ei teknologinen. Tarkastelen moraalisen vastuullisuuden käsitteellisiä ehtoja ja arvion niiden valossa, voisiko robotti kuulua vastuuasenteiden ja -käsitteiden alaan. Olen erityisen kiinnostunut siitä, onko olemassa jokin vastuullisuuden ehto, joka jää välttämättä täyttymättä robottien kohdalla. Antti Kauppinen kutsuu vastuupessimisteiksi filosofeja, joiden mukaan moraalisella vastuullisuudella on ehtoja, jotka eivät robottien tapauksessa täyty.[5] Tällaisiksi ehdoiksi katsotaan esimerkiksi tietoisuus tai vapaa tahto. Koska näyttää siltä, etteivät robotit kykene tahdonvapauteen eivätkä tietoisuuteen, ne eivät ansaitse osakseen vastuuasenteita. Vastuuoptimistit puolestaan ajattelevat, ettei meillä ole riittäviä perusteita sulkea pois sitä mahdollisuutta, että robotti voisi olla moraalisen yhteisömme jäsen ainakin joiltakin osin.

Artikkelini etenee seuraavasti. Esittelen ensin yleisiä näkökohtia moraaliseen vastuullisuuteen. Tämän jälkeen tarkastelen kahta pessimistien keskeistä argumenttia, joista ensimmäinen vetoaa robottitietoisuuden mahdottomuuteen

1 Ks. esimerkiksi Ranskan strategia Villani 2018.
2 Vrt. Ollila 2019, 224.
3 Robotin määrittelyn ongelmallisuudesta, ks. Coeckelbergh 2022.
4 Ks. esim. Nyholm 2020; Hakli & Seibt 2017. Hyödyllisen katsauksen tarjoaa Behdadi & Munthe 2020.
5 Kauppinen 2021, 142.

ja toinen taas robottien autonomian mahdottomuuteen. Tämän jälkeen tarkastelen joitakin perusteita varovaiselle optimismille robottien vastuullisuudesta. Viime aikoina suositut varovaisen optimistiset kannat esittävät, että vaikka robotti ei voisikaan olla täysimääräinen moraalinen toimija, se voisi ehkä silti olla sopiva kohde ainakin joillekin vastuuasenteille.

MORAALINEN VASTUULLISUUS JA JÄRKIPERUSTEHERKKYYS

On selvää, että moraalisen vastuullisuuden luonteesta ja sen ehdoista vallitsee filosofien keskuudessa merkittäviä erimielisyyksiä. Robottien mahdollinen moraalinen vastuullisuus riippuukin oikeastaan siitä, millainen näkemys moraalisesta vastuullisuudesta valitaan asian tarkastelemiseksi. Monet robottietiikan parissa toimivat filosofit ovat kuitenkin yhtä mieltä joistakin moraalisen vastuullisuuden käsitteellisistä ehdoista. Esittelen niitä seuraavaksi.

Ensimmäisenä voidaan kysyä, mikä moraalisen vastuullisuuden ilmiö oikein on.[6] Se, että henkilö A pitää henkilö B:tä vastuussa jostakin teosta, tarkoittaa sitä, että A suuntaa B:tä kohtaan tiettyjä asenteita ja muodostaa B:stä arvostelmia B:n tekojen perusteella. Filosofit kutsuvat tällaisia asenteita reaktiivisiksi asenteiksi, koska ne syntyvät psykologisesti vaivattomasti ja automaattisesti, osana ihmisten sosiaalista vuorovaikutusta.[7] Jos B on esimerkiksi jättänyt lupauksensa A:lle täyttämättä, A saattaa arvioida B:n epäluotettavaksi ihmiseksi, tulla vihaiseksi tätä kohtaan ja vaatia häntä tilille lupauksen rikkomisesta. Vastuussa pitäminen koostuu monenlaisista asenteista ja arvostelmista, vaikka filosofit usein keskittyvätkin negatiivisiin arvostelmiin ja asenteisiin, kuten esimerkiksi erilaisiin moitteen muotoihin sekä kaunaan.

Reaktiivisten asenteiden valossa moraalinen vastuullisuus on sekä psykologinen että sosiaalinen ilmiö. Ilmiö on psykologinen siinä mielessä, että se koskee moraalisia asenteita ja tunteita. Ilmiö on sosiaalinen siinä mielessä, että se koostuu ihmislajille tyypillisestä tavasta muodostaa yhteisöjä ja kuulua niihin. Vastuuasenteilla ja -käytännöillä, kuten syyttämisellä, katumuksella, rankaisemisella ja anteeksiannolla, on keskeinen rooli siinä, miten moraalinen yhteisömme toimii ja miten se kohtelee jäseniään.[8] Keskustelu robottien moraalisesta vastuullisuudesta koskee sitä, missä määrin on perusteltua laajentaa näitä asenteita ja käytäntöjä koskemaan myös robotteja. Keskeiseksi kysymykseksi nousee, missä määrin on sopivaa suhtautua robotteihin samalla tavalla kuin suhtaudumme moraalisen yhteisömme muihin jäseniin.

Tällainen lähestymistapa on siinä mielessä naturalistinen, että siinä moraalisen vastuullisuuden välttämättömiä ehtoja ei etsitä metafysiikasta, vaan ne johdetaan todellisuudessa esiintyvistä vastuuasenteista, normeista ja käytännöistä. On selvää, että reaktiivisten asenteiden sopivuutta säätelevät erilaiset normatiiviset odotukset. Vaikka henkilö B olisikin tehnyt väärin, ei tästä vielä seuraa se, että A:n häntä kohtaan omaksumat asenteet olisivat sopivia tai oikeutettuja. On siis olemassa normatiivisia ehtoja sille, milloin vastuuasenteet ovat sopivia ja milloin eivät. Voisimme kutsua näitä Peter Strawsonia seuraten vastuusta vapauttamisen ehdoiksi.

Ensimmäinen joukko ehtoja koskee sitä, onko B:tä sopivaa pitää laisinkaan henkilönä, joka voisi olla vastuuasenteiden kohteena. Esimerkiksi pienet lapset tai vakavasti sairaat tai kehitysvammaiset ihmiset katsotaan kyvyttömiksi kantamaan moraalista vastuuta. Jotkut ehdot koskevat henkilöitä, jotka katsotaan sopiviksi vastuuasenteiden kohteeksi yleisesti ottaen

6 Keskusteluni moraalisen vastuun luonteesta perustuu esitykseen Visala 2018, 50–63.

7 Reaktiivisten asenteiden käsite tulee Peter Strawsonin klassisesta artikkelista ”Freedom and Resentment”, joka löytyy kokoelmasta Strawson 2008. Ks. Strawson 2008, 6–7. Lyhyt esittely on Visala 2018, 55–60.

8 Näkökulmia moraalisen vastuullisuuden sosiaaliseen luonteeseen tarjoaa Hutchison et al. 2018.

(esimerkiksi normaalitilanteessa olevat aikuiset ihmiset) mutta jotka tulisi jossakin tilanteessa vapauttaa vastuusta. B voi kyetä moraalisesti vastuulliseen toimintaan mutta häntä ei ole syytä moittia väärästä teosta, koska hänellä on siitä vapauttava peruste. Esimerkiksi jos B on pakotettu tekemään väärin tai hän toimii tietämättään väärin, A:n olisi ehkä syytä pidättäytyä pitämästä B:tä vastuussa. On myös tapauksia, joissa A:n ei tule luopua kaikista asenteista, vaan hänen tulee hillitä tai säädellä niitä. Esimerkiksi jos B on vaikeasti masentunut ja hän pettää lupauksensa A:ta kohtaan, A:n tulisi tuskin pitää B:tä täydessä vastuussa lupauksensa rikkomisesta.

Kuten edelliset esimerkit paljastavat, vastuusta vapauttavia perusteita on käytännössä kahta lajia: tietämättömyys teon luonteesta ja vaikutuksista sekä kyvyttömyys kontrolloida kyseistä tekoa.[9] Kääntäen voidaan sanoa, että toimijaa on sopivaa pitää vastuussa sellaisista teoista, joiden moraalisen luonteen hän ymmärtää ja joita hän kykenee kontrolloimaan.

Ensimmäinen ehto on episteeminen eli tiedollinen: henkilön tulee tietää, mitä hän tekee moraalisessa mielessä.[10] Tämä koskee myös sitä, miksi henkilö tekee jonkin teon. Jos henkilö ei kykene tunnistamaan eikä ymmärtämään moraalisia vaatimuksia ja normeja, ei häntä tule silloin moittiakaan siitä, jos hän ei toimi moraalisten vaatimusten ja normien mukaisesti. Tämä ehto pätee myös reaktiivisten asenteiden ilmaisun mielekkyyteen. Ei näytä olevan mieltä ilmaista tyytymättömyyttä tai suuttumusta sellaista toimijaa kohtaan, joka ei ymmärrä mitä tällaiset asenteet edes ovat.

Sen lisäksi, että toimija tietää mitä hän tekee, on hänen kontrolloitava tekoaan. Filosofisessa kirjallisuudessa tätä kutsutaan usein vapaaksi tahdoksi (tai vastuun kontrolliehdoksi) ja sen olemassaolosta on mittava keskustelu.[11] En tässä yhteydessä halua puuttua tähän keskusteluun, vaan oletan että tekojen kontrollin tematiikkaa voidaan tarkastella siinä määrin pragmaattisesti, että on mahdollista erottaa ne teot, joita toimija itse kontrolloi teoista, joita hän ei kontrolloi (tai joiden osalta hänen kontrollinsa on alhainen). Esimerkiksi toimijan tietoisesta harkinnasta ja päätöksestä seuraava teko näyttää olevan paljon enemmän toimijan kontrollissa kuin teko, joka on tulosta esimerkiksi pakosta, riippuvuudesta tai manipulaatiosta. Mitä paremmin toimija kykenee reagoimaan erilaisiin toiminnan perusteisiin, muodostamaan toimintasuunnitelmia ja toteuttamaan aikomuksiaan, sitä paremmin hän kontrolloi tekojaan.

Edellisistä huomioista näemme, kuinka toimijat voidaan asettaa vastuullisuuden skaalalle. Skaalan alapäässä ovat toimijat, jotka eivät ylitä vastuullisen toimijan minimiehtoa, eli sitä, että toimija kykenee ottamaan huomioon reaktiivisia asenteita ja vastaamaan niihin käyttäytymisellään. Valtaosa filosofeista ajattelee, että tämä on mahdollista, jos toimija kykenee jonkinlaiseen järkiperusteherkkyyteen.[12] Vaikka toimija täyttäisikin kyseisen minimiehdon, ei se vielä tarkoita sitä, että häntä olisi sopivaa moittia kaikissa tilanteissa. On monia tekijöitä, jotka antavat perusteen pidättää tai hillitä reaktiivisia asenteita toimijan väärintekoa kohtaan. Esimerkiksi lievästä depressiosta kärsivää toimijaa ei ole ehkä sopivaa moittia tapaamisen unohtamisesta tai työtehtävissä myöhästymisestä. Skaalan yläpäässä ovat puolestaan toimijat, joita pidetään moraalisen yhteisön täysvaltaisina jäseninä ja joilla ei ole joissakin tilanteessa vastuusta vapauttavia erityisehtoja.

Teen jatkossa oletuksen, jonka jotkut filosofit kieltävät. Oletan, että moraalisen vastuun käytäntömme ja asenteemme ovat pääsääntöisesti oikeutettuja. Oletan siis, että olemme oikeutettuja pitämään täysivaltaisia moraali-

9 Strawsonia (2008, 7–8) mukaillen voisimme kutsua näitä ”en tiennyt” ja ”en voinut mitään” perusteiksi.

10 Ks. Robichaud & Wieland 2017.

11 Ks. Visala 2018.

12 Kaikkein tunnetuin teoria järkiperusteherkkyydestä on Fischer & Ravizza 1998. Ks. Visala 2018, 168–175.

sen yhteisömme jäseniä vastuussa teoistaan. Vapaata tahtoa ja moraalista vastuullisuutta koskevassa filosofisessa keskustelussa esiintyy nimittäin skeptisismiksi kutsuttu kanta, jonka mukaan moraalisen vastuullisuuden järjestelmämme ei ole oikeutettu ja sitä tulisi tästä syystä radikaalisti revisioida.[13] Jos tämä pitää paikkansa, myös robottien moraalinen vastuullisuus asettuu erilaiseen valoon. Jos meillä ei ole riittäviä perusteita pitää edes ihmisiä vastuussa teoistaan, tuskin on tarvetta pitää robottejakaan vastuullisina, jotta nämä voisivat kuulua moraaliseen yhteisöömme. Skeptikko joutuu kuitenkin vastaamaan itse peruskysymykseen tavalla tai toisella: voidaanko robotteja pitää yhteisömme jäseninä? Vaikka ajattelenkin, että skeptinen näkökulma on joiltakin osin perusteltu ja se voi opettaa meille paljonkin vapaudesta ja vastuusta, laitan sen jatkossa syrjään.

VASTUUPESSIMISMI JA TIETOISUUS

Vastuupessimistien arsenaalissa on monia argumentteja. Käsittelen seuraavaksi vain kahta niistä. Ensimmäinen koskee tietoisuutta. Esimerkiksi Brian Talbot, Ryan Jenkins ja Duncan Purves väittävät, ettei

> – – nykyisillä roboteilla ole fenomenaalista tietoisuutta, eikä sitä tule näillä lähitulevaisuudessa olemaakaan. Tästä syystä roboteilla ei ole niitä psykologisia kykyjä, joita toimijuus edellyttää. Valtaosa moraalista päätöksentekoa koskevista teorioista liittää sen fenomenaaliseen tietoisuuteen. – – On myös uskottavaa, että itsetietoisuus, moraalinen mielikuvitus ja intuitio edellyttävät sitä. – – Lisäksi, jos roboteilla ei ole fenomenaalista tietoisuutta, eivät ne myöskään pysty toimimaan järkiperusteiden pohjalta.[14]

Argumentin mukaan robotit eivät koskaan täytä vastuullisen toimijuuden minimivaatimusta, koska roboteilta ainakin nyt ja lähitulevaisuudessa puuttuu tietoisuus, joka on kaiken moraalisen toiminnan edellytys. Robotit eivät siis kykene ymmärtämään moraalisia näkökohtia eivätkä kontrolloimaan tekojaan sen nojalla.

Jotta argumentti olisi uskottava, olisi sen esittäjän seuraavaksi perusteltava kaksi väitettä. Ensinnäkin hänen olisi osoitettava, että moraalinen vastuullisuus todella edellyttää esimerkiksi kykyä moraaliseen päätöksentekoon, itsetietoisuuteen, mielikuvitukseen, intuitioon sekä järkiperusteiden pohjalta toimimiseen. Tämän jälkeen olisi osoitettava, että nämä kyvyt edellyttävät juuri tietynlaista tietoisuutta, fenomenaalista tietoisuutta.

Aloitetaan ensimmäisestä väitteestä. Kuten edellä totesin, lähden liikkeelle siitä, että järkiperusteherkkyys on minimiehto vastuullisuudelle. Ehkä täysissä ruumiin ja sielun voimissa toimivan ihmisyksilön moraalisiin kykyihin kuuluu esimerkiksi monimutkainen moraalinen mielikuvitus ja intuitio, mutta on vaikea pitää näitä vastuullisuuden välttämättöminä ehtona. Rima vastuulliselle toimijuudelle vaikuttaa olevan tässä aivan liian korkea. Osa sellaisistakin ihmisistä, joita tavallisesti pidetään vastuullisina toimijoina, putoaisi sen alapuolelle. Lisäksi näyttää siltä, että algoritmit kykenevät tekemään ei-moraalisia päätöksiä, joten ainakin päätöksenteko ylipäätään voidaan toteuttaa vailla fenomenaalista tietoisuutta. Keskityn siis jatkossa lähinnä kysymykseen järkiperusteherkkyydestä ja sen suhteesta tietoisuuteen.

Fenomenaalisella tietoisuudella viitataan henkilön laadullisiin kokemuksiin, eli siihen, miltä erilaiset kokemukset ”tuntuvat” hänen omasta näkökulmastaan.[15] Henkilön ollessa tietoinen esimerkiksi kivusta alaselässään hänellä on tietty kivun kokemus, joka ei ole havaittavissa kolmannen persoonan näkökulmasta. Fenome-

13 Skeptisismi on kasvattanut suosiotaan viime vuosina. Ks. Pereboom 2014, 2021; Caruso 2021; Waller 2015.

14 Talbot, Jenkins & Purves 2017, 258–259 (suomennos Aku Visala).

15 Hyvä yleiskatsaus tietoisuutta koskevaan filosofiseen väittelyyn on van Gulick 2021.

naalisen tietoisuuden olemassaolo ja kausaalinen rooli mielen toiminnassa ovat merkittäviä mielenfilosofian kiistakysymyksiä, joita en tässä yhteydessä voi käsitellä kattavasti.

Yksi tapa vastata pessimistin argumenttiin on hyvin suoraviivainen: kielletään fenomenaalisen tietoisuuden olemassaolo tai sen kausaalinen rooli mielen toiminnassa. Eliminativistit väittävät, että olemme yksinkertaisesti väärässä omista kokemuksistamme. Todellisuudessa meillä ei ole mitään ensimmäisen persoonan kokemuksia.[16] Toinen vaihtoehto on väittää reduktiivisten fysikalistien tavoin, että tietoisuuden kokemusominaisuudet ovat identtisiä esimerkiksi joidenkin aivotilojen tai informaatiotilojen kanssa. Jos fenomenaalinen tietoisuus edellyttää vain tiettyjen informaatiotilojen läsnäoloa järjestelmässä, ei näytä olevan mitään periaatteellista estettä sille, etteikö myös kone voisi olla tietoinen.

Esimerkiksi Daniel Dennett suhtautuu kriittisesti fenomenaalisen tietoisuuden olemassaoloon, mutta pitää siitä huolimatta ihmisiä pääsääntöisesti sopivina kohteina vastuuasenteille. Robottien osalta hän edustaa kantaa, jota kutsuin edellä varovaiseksi optimismiksi. Robotteja ei tule pitää vastuussa teoistaan. Syy tälle ei ole robottien puuttuva tietoisuus, vaan se, etteivät robotit vielä kykene muodostamaan käsityksiä omista (eivätkä toisten) mielentiloista.[17] Koska robotit ovat vielä tällä hetkellä kyvyttömiä tunnistamaan moraalisen toiminnan perusteita, niitä ei tule pitää vastuussa teoistaan. Mutta asiaan voi tulla muutos tulevaisuudessa.

Kauppinen kuvaa yhtä pessimistin argumenttia seuraavasti. Jotta toimija voi toimia moraalisilla perusteilla (ei siis pelkästään esimerkiksi oman edun tai hyödyn perusteella), tulee toimijan pystyä tekemään ero moraalisten ja ei-moraalisten perusteiden välillä. Tämä on mahdollista vain, jos henkilö välittää toisten kärsimyksestä ja voi kokea sen. Ilman fenomenaalista tietoisuutta toimijalla ei ole kokemuksia, joten tämä toimija ei voi välittääkään mistään, eikä näin ollen kykene erottamaan toisistaan moraalisia perusteita ei-moraalisista tekojen perusteista.[18]

On kuitenkin vaikea nähdä, miksi välittämiseen tarvittaisiin nimenomaan fenomenaalinen kokemus välittämisestä. John Martin Fischerin ja Mark Ravizzan teoria on luultavasti tunnetuin yritys kuvata sitä, mitä järkiperusteherkkyys on ja mitä se edellyttää.[19] Nähdäkseni Fischer ja Ravizza eivät missään kohdin vaadi, että toimijalla on fenomenaalinen tietoisuus. He kyllä edellyttävät toimijalta tietoisuutta.[20] Samassa hengessä olen itsekin väittänyt, että tietoisuus todella on moraalisesti vastuullisen toimijuuden välttämätön ehto. Ilman tietoisuutta toimija ei kykene käsittelemään moraalisia näkökohtia ja ohjaamaan tekojaan niiden perusteella.[21] Tähän vaadittu tietoisuus voidaan kuitenkin ymmärtää myös ilman fenomenaalista tietoisuutta.

Tahdonvapautta koskevassa filosofisessa ja tieteellisessä keskustelussa tietoisuudella on keskeinen asema. Kutsutaan nyt tietoisuusteesiksi sitä väitettä, että vastuullisuus edellyttää toimijalta tietoisuutta. Skeptikot kieltävät vapaan tahdon olemassaolon vetoamalla siihen, etteivät ihmiset ole pääsääntöisesti tietoisia omien tekojensa vaikuttimista. Tällöin ihmisten tietoisten aikomusten, päätösten ja päämäärien suhde heidän tekoihinsa on löyhä tai olematon. Näin ollen ihmisiä ei voi pitää vastuussa heidän teoistaan, koska teot eivät ole ihmisten tietoisten sisältöjen kontrollissa.[22] Tässä debatissa tietoisuus määritellään pääsytietoisuudeksi (engl. access consciousness) tai tilatietoisuudeksi (engl. state consciousness). Pääsytietoinen järjestelmä kykenee representoimaan oman itsensä ja osiensa toimintaa. Neil Levyn mukaan tietoisuusteesi täyttyy monien ihmisten

16 Ks. esim. Churchland 1994.
17 Dennett 1997.
18 Kauppinen 2021, 146.
19 Fischer & Ravizza 1998.
20 Levy 2014, 109.
21 Visala 2021.
22 Visala 2017 kuvaa skeptikkojen argumentaatiota.

kohdalla. Hän käyttää Bernard Baarsin ja hänen kollegoidensa teoriaa tietoisuudesta ”globaalina työtilana” (engl. global workspace theory).[23] Baarsin mukaan järjestelmä on tietoinen, kun sen tiedonkäsittelyllä on keskus, joka kerää syötettä alajärjestelmistä, yhdistelee sitä ja syöttää sen takaisin alajärjestelmille.[24]

Jos tietoisuusteesin oletetaan täyttyvän monien ihmisten osalta, en näe käsitteellistä estettä sille, etteikö robottien tiedonkäsittelyn arkkitehtuuri voisi noudattaa samaa mallia. Tätä väitettä tukee se empiirinen tosiseikka, että Baarsin teoriaa on käytetty tietynkaltaisten oppivien järjestelmien mallina jo pidemmän aikaa.[25] Näyttää siis siltä, että tietoisuusteesi on mahdollista pelastaa ilman fenomenaalisen tietoisuuden todistustaakkaa. Optimisti voi käyttää tätä argumentaatiolinjaa pessimistiä vastaan.

Vastuuoptimisti voi vedota myös siihen, etteivät spekulaatiot fenomenaalisesta tietoisuudesta ole tärkeitä arkielämän vastuukäytäntöjen kannalta. Ehkei ole merkitystä sillä, onko vastuullisella toimijalla jokin sisäinen kokemus, joka ei näy toiminnassa. Vastuun kannalta on merkityksellistä, millaisiin toimintoihin ja käyttäytymiseen toimija todella kykenee. Kysymys on siis siitä, missä määrin robotti kykenee todellisuudessa osallistumaan moraaliseen yhteisöömme eli ottamaan vastaan ja käsittelemään reaktiivisia asenteita. Wendell Wallach ja Colin Allen esittävät, että jäsenyyteen moraalisessa yhteisössä riittää funktionaalinen samankaltaisuus.[26] Heidän mukaansa on mahdollista, että keinotekoinen moraalinen toimija voi ainakin joiltakin osin osallistua siihen reaktiivisten asenteiden ”peliin”, josta vastuujärjestelmämme koostuu. Tällöin ei ole merkitystä sillä, mitä keinotekoisen toimijan ”sisällä” tapahtuu eli onko tällä toimijalla fenomenaalinen tietoisuus vai ei.

Wallachin ja Allenin lähestymistapa on suhteellisen tavallinen optimistien keskuudessa. Wulf ja Janina Loh kirjoittavat:

> Kysymys siitä, voiko keinotekoinen järjestelmä olla älykäs, tietoinen tai autonominen vahvassa mielessä, korvautuu kysymyksellä siitä, missä määrin järjestelmä täyttää ne funktiot, jotka ovat moraalisen arvioinnin kohteena, tässä tapauksessa moraalisen vastuullisuuden arvioinnin kohteena. Esimerkiksi keinotekoisen järjestelmän kykyä järkeillä tai toimia autonomisesti tarkastellaan vain siinä määrin kuin se toimii välttämättömänä ennakkoehtona sille, että järjestelmää pidetään vastuullisena.[27]

Näyttää siltä, että robotti voi olla järjestelmä, jonka kognitiivinen arkkitehtuuri on tietoinen edellä kuvatussa mielessä. Tällöin se voisi olla joiltakin osin funktionaalisesti samankaltainen niiden moraalisten toimijoiden kanssa, jotka jo nyt kuuluvat moraaliseen yhteisöömme. Kysymys on tästä eteenpäin lähinnä teknologinen, ei filosofinen: missä määrin kykenemme kehittämään järjestelmiä, jotka tunnistavat mahdollisimman laajan skaalan moraalisia perusteita ja kykenevät ohjaamaan toimintaansa niiden perusteella?

VASTUUPESSIMISMI JA AUTONOMIA

Wulf ja Janina Loh mainitsivat edellä autonomian käsitteen. Toinen tärkeä argumentti pessimistien patteristossa koskeekin juuri sitä. Argumentin mukaan vastuullisen kohtelun sopivuus ei riipu pelkästään siitä, millaisia kykyjä toimijalla on. Vastuullisuus edellyttää, että toimijalla on tietynkaltainen historia. Autonominen toimija on syntynyt juuri tietyllä tavalla – tavalla, johon ei kuulu ulkopuolinen, manipulatiivinen vaikutus. Argumentti etenee tästä väittämään, ettei robotti voi koskaan olla autonominen toimija, koska robotti on keinotekoinen, jonkun muun toimijan valmistama olio. Vaikka robotilla olisikin kyky

23 Levy 2014.
24 Baars 2002; Dehaene et al. 2011. Skeptistä näkemystä puolustaa esimerkiksi David Caruso (2012).
25 Esimerkiksi LIDA-arkkitehtuuri. Ks. https://ccrg.cs.memphis.edu.
26 Wallach & Allen 2009.
27 Loh & Loh 2017, 39.

tunnistaa järkiperusteita ja kontrolloida tekojaan niiden perusteella, siltä kuitenkin puuttuu vapaus manipulaatiosta eli autonomia.

Tarkastelen seuraavaksi lyhyesti Raul Haklin ja Pekka Mäkelän argumenttia. He kirjoittavat:

> Voi olla mahdollista ohjelmoida ja valmistaa robotti, jolla on kaikki moraaliseen toimijuuteen tarvittavat kyvyt, mutta tämä robotti ei vielä silti olisi moraalinen toimija. Tämä johtuu siitä, että toisin kuin aitojen moraalisten toimijoiden, joiden kyvyt on hankittu autenttisella tavalla, robottien kausaaliseen historiaan kuuluu välttämättä valmistusprosessi, joka lasketaan autonomian poissulkevaksi manipulaatioksi. Juuri se, että vastuullisuuden mahdollistavat kyvyt ja ominaisuudet ovat keinotekoisesti tuotettuja, tekee vastuuattribuutioista perusteettomia. Robotit eivät ole eivätkä koskaan voi olla sopivia kohteita moraalisessa vastuussa pitämiselle, koska toiset toimijat ovat suunnitelleet, rakentaneet ja ohjelmoineet niille sen ”luonteen”, joka niillä on.[28]

Haklin ja Mäkelän mukaan robotti ei voi koskaan olla vastuullinen toimija, koska tältä puuttuu sopiva kausaalinen historia. Robotin kausaalinen historia sisältää aina manipulaattorin, josta robotin päämäärät ja kyvyt riippuvat. Tämä manipulaattori on se ihminen, joka on robotin suunnitellut ja luonut. Robotti ei voi koskaan olla vastuuasenteiden kohde, koska joku muu on tehnyt sen tiettyä tarkoitusta varten. Tällöin vastuuasenteiden sopiva kohde on ainoastaan robotin suunnittelija ja tekijä.

On huomattavaa, ettei Haklin ja Mäkelän argumentti ole se, että robotit on ohjelmoitu toteuttamaan tekijänsä tahtoa. Heidän mukaansa on selvää, että robotit kehittyvät kohti yhä merkittävämpää autonomiaa siinä mielessä kuin käsitettä ”autonomia” käytetään teknologian kentällä. Tällöin autonomialla viitataan robotin kykyyn tehdä yhä itsenäisempiä päätöksiä yhä joustavammilla perusteilla. Tällaiseen autonomiaan jo jotkut nykyisetkin robotit kykenevät. Robotit eivät silti kykene autonomiaan sanan filosofisessa mielessä, koska niiden oppimisen ennakkoehdot ja käyttötarkoitus ovat robotin valmistajien määrittämiä.

Kuten arvata saattaa, kysymys autonomiasta ja moraalisesta vastuullisuudesta on kiistanalainen. Ensiksi on syytä tarkastella sen laajempaa taustaa. Autonomian välttämättömyyteen vedotaan silloin, kuin halutaan osoittaa determinismin ja vastuullisuuden yhteensopimattomuus. Manipulaatioargumentit on suunniteltu juuri tätä tehtävää varten.[29] Ne koostuvat pääsääntöisesti kahdesta osasta. Ensimmäinen väite on se, että jos toimijan A kyvyt, arvostukset ja päämäärät ovat tulosta toimijan B suunnitelmasta ja päämääristä, A:ta ei tule pitää vastuussa teoistaan. Tämä väite olettaa, että vastuullisuus on historiallinen käsite eli toimijan vastuullisuuteen ei riitä pelkästään se, millainen hän on teon hetkellä, vaan vastuullisilla toimijoilla on juuri tietynkaltainen historia. Toimija voi olla vastuussa teoistaan vain siinä tapauksessa, että hän on itse omien tekojensa lähde. Jos toimijan teot ovat tulosta päämääristä, luonteesta ja kyvyistä, jotka eivät ole toimijan omia, ei häntä voi pitää vastuussa teoistaan.

Manipulaatioargumentin toinen väite puolestaan on se, että jos determinismi metafyysisenä teesinä pitää paikkansa, on determinismillä sama vaikutus kaikkien toimijoiden historiaan kuin manipulaatiolla. Determinismin tapauksessa manipulaation lähteenä ei ole jokin toimija, vaan koko ”luonto” ja sen lait. Kummassakin tapauksessa toimija ei ole itse päämääriensä, kykyjensä ja taipumustensa lähde, vaan niiden lähde on toimijasta itsestään riippumaton.

Haklin ja Mäkelän käyttämä manipulaatioargumentti on lainattu Alfred Meleltä. Tämän argumentin tekee kiinnostavaksi se, että se nimenomaan ei pyri osoittamaan determinismin

28 Hakli & Mäkelä 2019, 269 (suomennos Aku Visala).

29 Manipulaatioargumenteista ks. Visala 2018, 101–109.

ja vastuun ristiriitaa. Sen sijaan Mele on vakuuttunut siitä, että vastuullisuus kyllä edellyttää tietynlaista kausaalista historiaa, mutta tämän kausaalisen historian ei tarvitse olla indeterminististen. Sen sijaan autonomiaan riittää se, ettei toimijalla ole historiaa, johon sisältyy toimijan kyvyt ja luonteen jotakin tarkoitusta varten suunnitellut manipulaattori.[30] Robottien tapauksessa tällainen manipulaattori on aina olemassa: se insinööri, joka robotin on suunnitellut. Tällöin olisi asiaankuuluvaa pitää vastuuasenteiden kohteena pikemminkin insinööriä kuin robottia.

Jos optimisti haluaa vastata Haklin ja Mäkelän argumenttiin, on hänellä nähdäkseni kolme vaihtoehtoa. Ensimmäiseksi hän voi kritisoida Haklin ja Mäkelän näkemystä autonomiasta. Vaikka pidänkin Haklin ja Mäkelän käyttämää Melen teoriaa autonomiasta hyvin perusteltuna, haluan tässä lähinnä huomauttaa, kuinka vaikean tehtävän edessä teoria on. On jonkin verran näyttöä siitä, että ihmiset todella pitävät autonomiaa vastuullisuuden välttämättömänä ehtona. Jos toimijan kyvyt, päämäärät ja taipumukset voidaan jäljittää jonkin toisen toimijan tahtoon ja toimintaan (”manipulaattori”), ihmiset tavallisesti pitävät manipuloitua toimijaa vähemmän vastuullisena.[31] Tätä ilmiötä kutsutaan manipulaatiointuitioksi. Manipulaatiointuition tulkinta ja autonomian puolustaminen sen perusteella ei kuitenkaan ole helppo tehtävä.

Haaste on se, että manipulaatiointuitio näyttää joko olevan liian vaativa tai liian vaatimaton. Jos vaadimme moraaliseen vastuullisuuteen liian vahvaa autonomiaa (esimerkiksi toimijan luonteen ja tekojen riippumattomuutta edeltävistä syistä), näyttää siltä, etteivät edes ihmiset kykene olemaan vastuullisia toimijoita. Autonomia lipuu kaikkien osapuolien saavuttamattomiin. Kun tarkastelemme moraaliseen yhteisöömme kuuluvien inhimillisten toimijoiden kausaalista historiaa, törmäämme aina tekijöihin, jotka ovat muokanneet toimijan moraalisia taipumuksia ja kykyjä mutta jotka kuitenkin ovat toimijasta itsestään riippumattomia. Emme voi juuri mitään kasvatuksellemme, geeneillemme ja kulttuurille, jossa kasvamme – puhumattakaan vanhempiemme, biologiamme ja kulttuurimme historiasta. Jos vastuullisuuden ehtona on riippumattomuus tällaisista vaikutuksista, vain harva ylittää riman. Jos taas rimaa lasketaan alaspäin ja robotit kehittyvät siinä mielessä autonomisemmiksi, että ne kykenevät muokkaamaan päämääriään ja oppimaan uutta, myös jotkut robotit saattavat päästä riman ylitse ja niitä voidaan pitää vastuullisina.

Optimistilla on myös muita strategioita käytettävänään. Kuten Hakli ja Mäkelä myöntävät, kaikki teoriat moraalisesta vastuullisuudesta eivät pidä autonomiaa vastuullisuuden välttämättömänä ehtona. Ehkä optimisti voisi hylätä autonomian ehdon kokonaan. Esimerkiksi attribuutioteoriat ja hierarkkiset teoriat eivät edellytä autonomiaa, vaan ne on nimenomaan suunniteltu toimimaan ilman sitä.[32] Näiden teorioiden mukaan vastuukäytäntömme ja asenteemme riippuvat vain siitä, millainen niiden kohteena oleva toimija on juuri nyt. Tällöin kriteerinä on lähinnä se, missä määrin toimija kykenee muodostamaan toisen kertaluokan arvioita omista haluistaan ja taipumuksistaan. Jos henkilön teot syntyvät haluista ja taipumuksista, jotka hän hyväksyy omikseen ja joiden ”takana hän seisoo”, on häntä kohtaan perusteltua suunnata reaktiivisia asenteita. Jos tällainen analyysi moraalisen vastuullisuuden ehdoista on oikea, omista arvoistaan ja päämääristään käsityksiä muodostava robotti joka on funktionaalisesti samankaltainen ihmisten kanssa voisi olla sopiva vastuuasenteiden kohde.

Kolmas mahdollinen strategia vastata Haklin ja Mäkelän argumenttiin on sen hyväksyminen osittain. Tällöin optimisti väittää, että vaikka manipulaatiosta vapaa historia olisikin täyden

30 Mele (1995) esittää teorian autonomiasta. Manipulaatiota ja autonomiaa käsittelee Mele 2019.

31 Ks. esim. Björnsson 2016. Mele (2019) pumppaa manipulaatiointuitioita esiin monin erilaisin esimerkein.

32 Talbert (2022) esittelee näitä teorioita. Ks. myös Visala 2018, 157–162.

moraalisen vastuullisuuden välttämätön ehto, se ei ehkä ole välttämätön ehto kaikkien vastuuasenteiden sopivuudelle. Voi hyvinkin olla, että autonomia on välttämätön ehto sille, että toimijaa voidaan rangaista lain edessä tai että häntä voidaan vaatia julkisesti puolustamaan itseään. Mutta ehkä joitakin reaktiivisia asenteita, kuten katkeruutta, kiitollisuutta ja kunnioitusta, voidaan perustellusti suunnata myös vähemmän autonomisia toimijoita kohtaan. Tarkastelen seuraavaksi tätä mahdollisuutta tarkemmin.

VAATIMATON OPTIMISMI JA OSITTAINEN VASTUULLISUUS

Olen nyt tarkastellut kahta pessimistien keskeistä argumenttia ja esittänyt muutaman tavan, jolla ne voitaisiin kiertää. Filosofisessa keskustelussa pessimistinen kanta on ollut valtavirtaa, kun taas optimistit ovat olleet vähemmistössä. Aivan viime vuosina varovaiselle optimistiselle kannalle on ilmestynyt yhä enemmän puolustajia.[33] Optimistit ovat kuitenkin varovaisia siinä mielessä, että he usein pitävät robotteja sopivina kohteina vain joillekin vastuuasenteille.

Dane Gogoshin on puolustanut varovaista optimismia robottien vastuullisuudesta.[34] Kuten edellä kävi ilmi, tavallisesti robottieetikot pitävät järkiperusteherkkyyttä moraalisen vastuullisuuden minimiehtona. Gogoshin haastaa tämän oletuksen. Hänen mukaansa moraalisen vastuun käytäntömme eivät pyri muokkaamaan sitä, mitä toimijan psykologiassa tapahtuu, vaan niiden tarkoituksena on muokata toimijoiden käyttäytymistä. Reaktiiviset asenteet ovat ihmisten tapa muokata toistensa käyttäytymistä. Gogoshinin mukaan vastuuasenteemme edellyttävät ”herkkyyttä, ei niinkään moraalisille toiminnan perusteille, vaan sosiaalisesta hyväksynnästä ja paheksunnasta syntyvälle nautinnolle ja kivulle”.[35] Toimijan ei siis tarvitse kyetä kognitiivisesti prosessoimaan moraalisia perusteita, vaan olemaan sellaisen vaikutuksen kohteena, joka vahvistaa moraalisesti suotavaa käyttäytymistä ja karsii moraalisesti epäsuotuisaa käyttäytymistä. Moraalisessa vastuussa pitämisen asenteet ja tunteet muodostavat sosiaalisen järjestelmän, joka muokkaa ihmisten käyttäytymistä kohti sosiaalisesti hyväksyttyjä normeja. Tämän järjestelmän olemassaolo oikeutetaan eteenpäin katsovilla perusteilla, eikä esimerkiksi sillä, millaista kohtelua ihmiset tosiasiassa ansaitsisivat.[36]

Gogoshinin tapa analysoida vastuujärjestelmäämme on yhteensopiva edellä mainitun funktionaalisen samankaltaisuuden ehdon kanssa. Vastuujärjestelmä ei tämän näkemyksen mukaan edellytä jotain tiettyä kognitiivista prosessia, joka tapahtuu toimijan pään sisällä. Riittää, että toimija kykenee reagoimaan reaktiivisiin asenteisiin muuttamalla käyttäytymistään, tai kuten Gogoshin asian ilmaisee, ”tuntee reaktiivisten asenteiden piston” ja päivittää käyttäytymistään sen perusteella.

Gogoshin väittää myös, että

> – – moraalinen toimijuus asettuu skaalalle. Skaalan korkeimmassa päässä on moraalinen autonomia. Jossakin kohtaa skaalaa saavutamme moraalisen kompetenssin kynnyksen, jonka jälkeen olemme moraalisessa vastuussa. Ehdotan, että tämä kynnys on kyky toimia luotettavalla tavalla moraalisten normien mukaan. – – Väitän, että moraalisia normeja seuraavat robotit, joilla on kyky täyttää sosiaalisia rooleja koskevat normatiiviset odotukset, ovat täten vastuussa teoistaan.[37]

Kuten lainauksesta käy ilmi, Gogoshin pitää moraalista autonomiaa arvokkaana. Se ei kuitenkaan voi olla moraalisen vastuullisuuden

33 Ks. List 2021; Tigard 2021.
34 Gogoshin 2021.
35 Gogoshin 2021, 5 (suomennos Aku Visala).
36 Gogoshin perustelee vastuujärjestelmän olemassaolon instrumentaalisesti (2021, 5). Sen olemassaolo on perusteltua siksi, että se muokkaa yksilöiden käyttäytymistä kohti muotoja, jotka ovat yhteisesti hyväksyttyjä ja hyödyllisiä myös yksilöille itselleen. Näiltä osin Gogoshin on hyvin lähellä vastuurevisionismia. Ks. Vargas 2013.
37 Gogoshin 2021, 7 (suomennos Aku Visala).

välttämätön ehto. Vain harva ihminen saavuttaa sen ja vastuuasenteemme eivät tähtää sen saavuttamiseen. Sen sijaan vastuuasenteidemme tarkoitus on se, että niiden avulla saamme toiset (ja itsemme) käyttäytymään moraalisten normien mukaisesti. Näin ollen, jos toimijalla on kyky ”tuntea reaktiivisten asenteiden pisto” ja tätä kautta muokata käyttäytymistään moraalisten odotusten mukaiseksi, on tämä toimija ylittänyt moraalisen vastuullisuuden minimikynnyksen.[38] Jos tämä on oikea analyysi vastuujärjestelmästämme ja sen oikeutuksesta, ei näytä olevan estettä sille, että myös itseohjautuva robotti voisi olla sopiva vastuuasenteittemme kohde.

Kuten myös Gogoshin itse tunnustaa, hänen analyysinsa on yleisluontoinen siinä mielessä, ettei se erittele tarkemmin sitä, millaisten reaktiivisten asenteiden kohteena robotti voisi perustellusti olla. Ei hänkään ajattele, että vastuun minimiehdon täyttäminen johtaisi täyteen vastuullisuuteen. Jotta varovainen optimisti voisi puolustaa kantaansa, olisi hyvä, jos hän pystyisi erottelemaan hieman tarkemmin ne reaktiiviset asenteet, jotka voisivat olla sopivia.

Vastuupluralistit ajattelevat, etteivät kaikki reaktiiviset asenteet ole samanlaisia.[39] Vastuupluralistien mukaan reaktiiviset asenteet voidaan jakaa ainakin kolmeen eri ryhmään. Kunkin ryhmän kohde ja tarkoitus on hieman erilainen. Samoin kullakin ryhmällä voi olla toisistaan poikkeavat sopivuusehdot. Reaktiiviset asenteet syntyvät tämän näkemyksen mukaan suhteessa toimijuuden eri ulottuvuuksiin. David Shoemakerin mukaan näitä on kolme: toimijan moraalinen luonne, toimijan moraalinen arviointikyky sekä toimijan kyky ottaa huomioon muiden intressit ja vaatimukset. Tällainen ”kolminainen” teoria vastuullisuudesta antaa vaatimattomalle optimistille mahdollisuuden väittää, että vaikka jotkut vastuuasenteet robotteja kohtaan eivät olisikaan sopivia, ehkä jotkut toiset ovat. [40] Esittelen jatkossa vastuupluralismia Shoemakerin pohjalta.[41]

Ensimmäistä vastuussa pitämisen muotoa voidaan kutsua syyksilukemiseksi (engl. attributability).[42] Syyksilukeminen tarkoittaa sitä, että toimijan tekoja pidetään hänen arvojensa, päämääriensä ja aikomustensa – hänen minuutensa ja tahtonsa – ilmauksina. Tällöin vastuullisen ja vastuuttoman toimijan ero on siinä, missä määrin toimijan teko ilmaisee hänen moraalisia taipumuksiaan. Esimerkiksi jos henkilö pakon uhalla saadaan ryöstämään pankki, ryöstö ei ilmaise hänen todellista luonnettaan, joten sitä ei tulisi lukea hänelle syyksi. Tyypillisiä syyksilukemisasenteita ovat esimerkiksi halveksunta, viha, arvonanto ja kunnioitus.

Toinen vastuun muoto kohdistuu Shoemakerin mukaan toimijan kykyyn ajatella moraalisesti (engl. answerability). Tällöin vastuuasenteiden kohteena ei ole toimijan luonne, vaan hänen kykynsä tehdä moraalisia arvostelmia eli vetää johtopäätöksiä moraalisista perusteista. Jos henkilö ryöstää pankin omasta mielestään hyvin perustein, eivätkä nuo perusteet kestä päivänvaloa, voidaan häntä pitää vastuussa odotukset alittavasta moraalisesta päättelystä. Tällöin on perusteltua pitää väärintekijää negatiivisten sanktioiden kohteena. Nämä sanktiot, vaikkapa moittiminen ja pettymys, olisivat epäreiluja tilanteessa, jossa henkilö ei kykenisi tekemään toisin kuin hän teki. Negatiivisten reaktiivisten asenteiden ilmaisulla on siis väärintekijää ohjaa-

38 Vaikka tämä ehto vaikuttaa vaatimattomalta, se ainakin pintapuolisesti näyttää edellyttävän fenomenaalista tietoisuutta. Miten robotti voisi ”tuntea reaktiivisten asenteiden piston” ilman fenomenaalista kokemusta? Kiitän anonyymiä vertaisarvioijaa tästä huomiosta. Ongelman ratkaisemiseksi olisi esitettävä analyysi reaktiivisten asenteiden kohteena olemisen vaikutuksista, joka ei edellyttäisi fenomenaalista tietoisuutta. Analyysi on kuitenkin liian mittava tehtävä tähän artikkeliin.

39 Jeppson 2022.

40 Ks. esim. Tigard 2021.

41 Shoemaker 2011, 2015.

42 Käsittääkseni vakiintunutta suomenkielistä terminologiaa ei ole. Käytän tässä yhteydessä käsitteitä samalla tavalla kuin Visala 2018, 52.

va tehtävä: ne vaativat häneltä parempaa moraalista järkeilyä jatkossa. Tyypillisiä asenteita tässä yhteydessä ovat ylpeys, hyväksyntä, arvonanto, pettymys sekä paheksunta.

Kolmatta vastuullisuuden muotoa voitaisiin ehkä kutsua suomeksi tilivelvollisuudeksi (engl. accountability).[43] Tällä viitataan puolestaan vaatimukseen, joka voidaan hyvin perustein asettaa väärintekijän kannettavaksi. Tällaiset vastuuasenteet suuntautuvat siihen, missä määrin toimija ottaa toisten toimijoiden intressit huomioon. Shoemaker mainitsee tilivelvollisuusasenteina esimerkiksi kiitollisuuden, syyllisyyden, loukkaantumisen ja katkeruuden asenteet.

Shoemakerin teorian tarkoituksena on selittää, miksi vastuuasenteemme erilaisia ”marginaalisia toimijoita” kohtaan ovat ambivalentteja. Shoemaker tarkastelee esimerkiksi autismin kirjolla olevia toimijoita, psykopatiaa ja depressiota.[44] Vaikkei Shoemaker tarkastelekaan robotteja, hänen teoriaansa voidaan silti soveltaa myös keinotekoisten olioiden tapaukseen. Niitä voitaisiin pitää jonkinlaisia marginaalitoimijoina.

Syyksilukeminen edellyttää kykyä toimia moraalisissa yhteyksissä tavalla, joka ilmentää jonkinlaista koherenttia kaavaa. Toimija siis kykenee järjestelmällisesti reagoimaan toiminnallaan erilaisiin moraalisiin normeihin ja haasteisiin. Jos toimija toimii täysin sattumanvaraisesti, yhdessä tilanteessa normien mukaisesti oikein ja toisessa vastaavassa taas väärin, ei toimijalla ole moraalista luonnetta laisinkaan. Robotiikan edistyessä on mielestäni mahdollista, että robotit kykenisivät yhä paremmin vastaamaan moraalisiin haasteisiin säännönmukaisilla tavoilla. Tällöin voisi olla sopivaa muodostaa arvostelmia siitä, onko jokin robotti moraalisesti hyvä vai paha tai omaksua arvonannon tai vihan asenteita robottia kohtaan. Ajattelen siis, että ainakin jotkut syyksilukemisen muodot voisivat olla sopivia robotteja kohtaan. Tämä näkemys sopii hyvin yhteen myös Gogoshinin arvion kanssa.

Tilivelvollisuus ja moraalinen päättely ovat puolestaan hankalampia roboteille. Näiltä osin toiminnallinen vastaavuus ihmisten ja robottien välillä on vielä hyvin kaukana. Nykyteknologian valossa keinotekoinen moraalinen päättely ei ole vielä robotiikan näköpiirissä. Roboteilla on vielä hyvin rajalliset kyvyt ilmaista tunteita tai tunnistaa ihmisten odotuksia, tunteita ja intressejä. Robotit eivät ole siis tarpeeksi sensitiivisiä moraalisille signaaleille voidakseen olla vastuullisia toimijoita. Toki koneoppimisen avulla toimivien robottien lisääntyvä autonomia voi edesauttaa moraalisten signaalien erottelukyvyn kehittymistä. Myös emotionaalisesti sensitiivinen robotiikka on edistynyt viime vuosina. Samoin laajoista aineistoista oppiminen voi parantaa robottien kykyä tunnistaa sosiaalisia ja moraalisia odotuksia. Haasteen tilivelvollisuuden toteutumiselle aiheuttaa kuitenkin robottien vaikeudet olla dialogissa ihmisten kanssa. Merkittävä osa vaikkapa moitteesta ja paheksunnasta on muodoltaan kielellistä. Tilivelvollisuusvastuu on vaatimus osallistua eräänlaiseen dialogiin, jossa kaksi toimijaa ilmaisee asenteitaan (kielellisesti ja ei-kielellisesti) ja vastaa niihin. Kyky dialogin keinotekoiseen tuottamiseen ei kuitenkaan vaikuta mahdottomalta vaatimukselta, kuten esimerkiksi Chat GPT-4 osoittaa.

ROBOTIT JA MORAALISEN YHTEISÖN DILEMMA

Haluan lopuksi nostaa esiin robottietiikkaa ja vastuullisuutta koskevan käytännöllisen dilemman.[45] Kuten edellä kävi ilmi, kysymys robottien moraalisesta vastuullisuudesta on kysymys siitä, missä määrin robotteja tulisi pitää sopivina kohteina vastuukäytännöillemme. Dilemma syntyy siten, että meillä näyttää olevan hyviä syitä laa-

43 Gary Watson teki erottelun tilivelvollisuuden ja syyksilukemisen välillä artikkelissaan ”Two Faces of Responsibility”, joka löytyy kokoelmasta Watson 2004.

44 Shoemaker 2015.

45 Dilemma voidaan jäljittää Immanuel Kantin (ja ehkä Tuomas Akvinolaisen) eläinten moraalista kohtelua koskeviin näkemyksiin. Kiitän anonyymiä vertaisarvioijaa tästä huomiosta. Ks. esim Darling 2021 ja Visala 2021.

jentaa moraalisen yhteisömme piiriä koskemaan myös robotteja, mutta myös hyviä syitä olla tekemättä näin.

Dilemman ensimmäinen sarvi on seuraava. Meillä on hyviä perusteita olla varovaisia sen suhteen, missä määrin otamme robotteja moraalisen yhteisömme jäseniksi. Syy on se, että yhteisön jäsenet muokkaavat niitä normatiivisia odotuksia, asenteita ja käsitteitä, joita yhteisö käyttää. Moraaliset käsitteemme ja asenteemme ovat elintärkeitä yhteisömme toiminnan ja olemassaolon kannalta. Jos otamme kyvyiltään ja historialtaan toisenlaiset oliot yhteisömme jäseniksi, vaikkapa anteeksiannon, moitteen, rakkauden ja rankaisemisen käytännöt ja asenteet voivat trivialisoitua ja ohentua. Jos alamme esimerkiksi kohdella ihmisen ja seksirobotin suhdetta ikään kuin se olisi rakkaussuhde, on vaarana, että se rakkauden käsite, joka viittaa kahden tasavertaisen toimijan emotionaaliseen ja tahdonalaiseen suhteeseen, kapenee tai häviää.[46] Seksirobotti ei ole moraalinen toimija eikä kykene tasavertaisen rakkaussuhteen edellyttämään autonomiaan.

Dilemman toinen sarvi on seuraava. Jos taas kieltäydymme laajentamasta moraalisen yhteisömme piiriä, vaikka robotit kehittyisivätkin yhä itseohjautuvammiksi ja sensitiivisemmäksi moraalisten odotusten suhteen, voi tämä tuottaa haittaa meillä itsellemme ja tehdä robottien kehittymisestä vaikeampaa. Kuten edellä kävi ilmi, koko vastuujärjestelmämme olemassaoloa voidaan perustella vetoamalla sen hyötyihin. Nämä hyödyt koskevat sekä niitä toimijoita, jotka pitävät toisia vastuussa, että vastuussa pidettyjä itseään.

Ajatellaan marginaalisia moraalisen yhteisömme jäseniä. Valtaosa ihmisistä myöntää, ettei pieni lapsi ansaitse moitetta ja rangaistusta hänen moraalisesti vääristä teoistaan – ainakaan samassa mittakaavassa kuin aikuinen ihminen. Tämän seikan voisi perustella vetoamalla siihen, että lapsen on usein vaikea ymmärtää moraalisia näkökohtia ja kontrolloida tekojaan. Silti pidämme sopivana, että lapsia kohtaan suunnataan kevennettyjä tai ”simuloituja” vastuuasenteita. Jos näin ei toimittaisi, lapset eivät koskaan oppisi yhteisön normeja ja odotuksia. Jotta toimijasta voi tulla moraalisesti vastuullinen toimija, on hänen kasvettava ja tultava osaksi moraalista yhteisöä. Vastuulliseksi toimijaksi kehittyminen edellyttää siis sekä moraalista yhteisöä ja sitä, että tuo yhteisö kohtelee kehittyvää toimijaa ”ikään kuin” tämä olisi vastuullinen toimija. Näin voitaisiin väittää, että jos suljemme itseohjautuvat ja sosiaaliset robotit vastuuasenteiden ulkopuolelle vain siksi, että ne eivät ole fenomenaalisesti tietoisia tai autonomisia, riistämme heiltä mahdollisuuden oppia sosiaalisia normeja ja päivittää käyttäytymistään.

Tähän voidaan vielä lisätä toinen huomio. Välinpitämätön suhtautuminen toisiin olioihin saattaa vahingoittaa toimijaa itseään. Vaikka robotti ei olisikaan tietoinen eikä kykenisi kärsimään, ihmisillä on kuitenkin vahva kognitiivinen taipumus ihmistää robotteja.[47] Tämä ihmistämisen taipumus tekee robottien kohtelusta moraalisesti relevanttia ihmisille. Esimerkiksi jos ihminen systemaattisesti kohtelee sosiaalista robottia kaltoin, tämä voi edesauttaa julmuuden ja kovasydämisyyden paheiden kehittymistä. Tällöin voi hyvinkin olla niin, että ihmisten pyrkimys kultivoida moraalisesti arvokkaita taipumuksia ja kykyjä, kuten empatiaa ja toisten huomioonottamista, voi edellyttää, että myös robotteja kohdellaan tietyllä tavalla.[48]

LOPUKSI

Olen tässä artikkelissa esitellyt viimeaikaista keskustelua robottien mahdollisesta moraalisesta vastuullisuudesta. Käsittelin ensin kahta vastuupessimistien keskeistä argumenttia, joista ensimmäinen vetosi robottien tietoisuuden puutteeseen ja toinen robottien puutteelliseen

46 Turkle 2007, 100–101. Ks. myös Ollila 2019, 200–212.

47 Laakasuo, Visala & Palomäki 2020, 138–139. Termi ”ihmistäminen”, Ollila 2019, 244.

48 Ks. Darling 2021.

autonomiaan. Esitin joitakin kriittisiä huomioita näistä argumenteista ja hahmottelin sen jälkeen varovaista optimismia koskien robottien vastuullisuutta. Varovaisen optimismin mukaan jotkut vastuuasenteet voivat olla sopivia robotteja kohtaan, mikäli robotit saavuttavat funktionaalisen samankaltaisuuden ihmisen kanssa. Varovaisen optimistin argumentti lepää tietynlaisen vastuuasenteittemme ja niiden tosiasiallisen oikeuttamista koskevan teorian varassa. Lopuksi mainitsin käytännöllisen dilemman, johon robottieettikko törmää riippumatta siitä, mitä hän ajattelee robottien mahdollisesta vastuullisuudesta.

Filosofinen keskustelu robottien mahdollisesta moraalisesta vastuullisuudesta opettaa meille sen, että kysymys on lopulta siitä, miten analysoimme ja perustelemme moraalisen vastuullisuuden järjestelmämme kokonaisuudessaan. Asetammeko vastuullisuudelle korkeat ehdot, kuten vaikkapa vahvasti metafyysisesti latautuneen käsityksen tahdonvapaudesta tai autonomiasta, vai riittääkö vähemmän vaativa ehto, kuten ”reaktiivisten asenteiden pisto”? Tuleeko vastuujärjestelmä perustella viime kädessä oikeudenmukaisuuteen ja reiluuteen koskevilla, taaksepäin katsovilla perusteilla, vai välineellisillä, eteenpäin katsovilla perusteilla? Nämä kysymykset ovat keskiössä nimenomaan tahdonvapautta ja moraalista vastuullisuutta koskevan filosofian piirissä. Robottietiikka näyttäytyy eri valossa riippuen siitä, millaisiin johtopäätöksiin tässä debatissa päädytään.

TT Aku Visala on uskonnonfilosofian dosentti ja yliopistotutkija Helsingin yliopistossa. Hänen nykyinen tutkimusprojektinsa käsittelee tahdonvapautta ja moraalista vastuullisuutta. Visala on kiinnostunut teologian, filosofian ja kognitiotieteen risteyskohdista.

• aku.visala@helsinki.fi

KIRJALLISUUS

Baars, Bernard (2002). The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Sciences* 6, 47–52.

Behdadi, Dorna & Christian Munthe (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines* 30, 195–218.

Björnsson, Gunnar (2016). Outsourcing the Deep Self: Deep Self Discordance Does Not Explain Away Intuitions in Manipulation Arguments. *Philosophical Psychology* 29, 637–653.

Caruso, Gregg (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lanhan: Lexington Books.

Caruso, Gregg (2021). *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Cambridge: Cambridge University Press.

Churchland, Patricia (1994). Can Neurobiology Teach Us Anything about Consciousness? *Proceeding and Addresses of the American Philosophical Association* 67:4, 23–40.

Coeckelbergh, Mark (2022). *Robot Ethics*. Cambridge, MA: The MIT Press.

Darling, Kate (2021). *The New Breed: What Our History with Animals Reveals about Our Future with Robots*. New York: Henry Holt & Co.

Dehaene, Stanislas, Jean-Pierre Changeux & Lionel Naccache (2011). The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architecture to Clinical Applications. *Characterizing Consciousness: From Cognition to the Clinic*. Toim. Stanislas Dehaene & Y. Christen. Berlin: Springer, 55–84.

Dennett, Daniel (1997). When HAL Kills, Who's to Blame? Computer Ethics. *HAL's Legacy: 2001's Computer as Dream and Reality*. Toim. D. G Stork. Cambridge, MA: MIT Press, 351–365.

Fischer, John Martin & Mark Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gogoshin, Dane Leigh (2021). Robot Responsibility and Moral Community. *Frontiers in Robotics and AI,* 8/768092. https://doi.org/10.3389/frobt.2021.768092.

Hakli, Raul & Pekka Mäkelä (2019). Moral Responsibility of Robots and Hybrid Agents. *Monist* 102, 259–275. doi:10.1093/monist/onz009.

Hakli, Raul & Johanna Seibt, toim. (2017). *Sociality and Normativity for Robots: Philosophical Inquiries into Human-Robot Interaction*. Dordrecht: Springer.

Hutchison, Katrina, Catriona MacKenzie & Marina Oshana, toim. (2018). *Social Dimensions of Moral Responibility*. Oxford: Oxford University Press.

Jeppson, Sofia (2022). Accountability, Answerability, and Attributability: On Different Kinds of Moral Responsibility. *Oxford Handbook of Moral Responsibility*. Toim. Dana Kay Nelkin & Derk Pereboom. Oxford: Oxford University Press, 73–89.

Kauppinen, Antti (2021). Osaammeko rakentaa moraalisia toimijoita? *Tekoäly, ihminen ja yhteiskunta: filosofisia näkökulmia*. Toim. Panu Raatikainen. Helsinki: Gaudeamus, 129–154.

Laakasuo, Michael, Aku Visala & Jussi Palomäki (2020). Kuinka ihmismieli vääristää keskustelua tekoälyn riskeistä ja etiikasta? Kognitiotieteellisiä näkökulmia keskusteluun. *Ajatus* 77:1, 131–167.

Levy, Neil (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

List, Christian (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology* 34, 1213–1242. doi:10.1007/s13347-021-00454-7.

Loh, Wulf & Janina Loh (2017). Autonomy and Responsibility in Hybrid Systems. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Toim. Patrick Lin, Ryan Jenkins & Keith Abney. Oxford: Oxford University Press, 35–50.

Mele, Alfred (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford: Oxford University Press.

Mele, Alfred (2019). *Manipulated Agents: A Window to Moral Responsibility*. Oxford: Oxford University Press.

Nyholm, Sven (2020). *Humans and Robots: Ethics, Agency, Anthropomorphism*. London: Rowman & Littlefield.

Ollila, Maija-Riitta (2019). *Tekoälyn etiikkaa*. Helsinki: Otava.

Pereboom, Derk (2014). *Free Will, Agency, and the Meaning of Life*. Oxford: Oxford University Press.

Pereboom, Derk (2021). *Wrongdoing and the Moral Emotions*. Oxford: Oxford University Press.

Robichaud, Philip & Jan Willem Wieland, toim. (2017). *Responsibility: The Epistemic Condition.* Oxford: Oxford University Press.

Talbert, Matthew (2022). Attributionist Theories of Moral Responsibility. *Oxford Handbook of Moral Responsibility*. Toim. Derk Pereboom & Dana Kay Nelkin Oxford: Oxford University Press, 53–70.

Talbot, Brian, Ryan Jenkins & Duncan Purves (2017). When Robots Should Do the Wrong Thing. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Toim. Patrick Lin, Ryan Jenkins & Keith Abney. Oxford: Oxford University Press, 258–273.

Tigard, Daniel (2021). Artificial Moral Responsibility: How We Can and Cannot Hold Machines Responsible. *Cambridge Quarterly of Healthcare Ethics* 30, 435–447. doi:10.1017/S0963180120000985.

Turkle, Sherry (2007). *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York: Basic Book.

Shoemaker, David (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics* 121:3, 602–632.

Shoemaker, David (2015). *Responsibility from the Margins*. Oxford: Oxford University Press.

Strawson, Peter (2008). *Freedom and Resentment and Other Essays*. London: Routledge.

Van Gulick, Robert (2021). Consciousness. *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). Toim. Edward N. Zalta. https://plato.stanford.edu/archives/win2021/entries/consciousness/.

Vargas, Manuel (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Villani, Cedric (2018). *For a Meaningful Artificial Intelligence: Towards a French and European Strategy*. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

Visala, Aku (2017). Neuronitko syypäitä? Viimeaikaista keskustelua vapaasta tahdosta ja neurotieteestä. *Ajatus* 74:1, 41–82.

Visala, Aku (2018). *Vapaan tahdon filosofia*. Helsinki: Gaudeamus.

Visala, Aku (2021). Consciousness and Moral Responsibility: Skeptical Challenges and Theological Reflections. *Zygon* 56:3, 641–665.

Visala, Aku (2021). Moraalinen toimijuus ja ihmiskeskeisyyden dilemma tekoälyn maailmassa. *Tekoäly, ihminen ja yhteiskunta: Filosofisia näkökulmia*. Toim. Panu Raatikainen Helsinki: Gaudeamus, 155–178.

Wallach, Wendell & Colin Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Waller, Bruce (2015). *The Stubborn System of Moral Responsibility*. Cambridge, MA: The MIT Press.

Watson, Gary (2004). *Agency and Answerability: Selected Essays*. Oxford: Oxford University Press.