Osaammeko rakentaa moraalisia toimijoita?

Antti Kauppinen (antti.kauppinen@helsinki.fi)

Helsingin yliopisto

Luonnos teosta *Tekoäly, ihminen ja yhteiskunta* varten, 14.3.2019

Sinä ja minä olemme moraalitoimijoita. Karkeasti sanottuna, kykenemme siis huomioimaan moraalisia ja moraalisesti relevantteja seikkoja päättäessämme mitä tehdä. Jos meillä ei olisi tällaista kykyä, toimisimme vielä useammin väärin, eikä meitä voisi edes pitää siitä täysin vastuullisena. Moraalitoimijuus on siis tärkeää. Mutta mitä se tarkalleen vaatii? Paul Grice esitti 70-luvun puolivälissä yhden vaikutusvaltaisen menetelmän toimijuuden olennaisten piirteiden tunnistamiseksi. Hänen ideansa oli, että ymmärtääksemme toimijuuden ehtoja, voimme

> rakentaa (tietysti vain mielikuvituksessa) tiettyjen periaatteiden mukaisesti tietynlaisen olennon, tai oikeastaan sarjan olentoja [yksinkertaisesta monimutkaiseen], jotka voivat toimia mallina tosiasiallisille olennoille. (Grice 1974-75, 37)

Grice ehdotti, että mielikuvitusolentojen ajatteleminen voi olla hyödyllistä, jos haluamme ymmärtää, mitkä ovat toimijuuden olennaiset rakennuspalikat. Mutta miksi tyytyä mielikuvitukseen? Jos voisimme aidosti rakentaa moraalitoimijoita, ymmärtäisimme itseämme aivan uudesta perspektiivistä. 1970-luvulla sellainen olisi ollut silkkaa scifiä. Mutta nyt olemme oppineet mallintamaan älykkyyttä vaativaa toimintaa ja hahmottamista ennennäkemättömällä tavalla. Vaikuttaa todennäköiseltä, että varsin läheisessä tulevaisuudessa voimme delegoida monien tarpeidemme tyydyttämisen koneille, jotka hoivaavat, hellivät, kuljettavat ja jopa puolustavat meitä fiksusti ja palautteesta oppien. Samaan aikaan pilvipalveluissa piilottelevat algoritmit tekevät päätöksiä luotoista ja

avustuksista ja valikoivat meille uutisia ja mainoksia. On luonnollista ajatella, että kun toimimme enenevässä määrin vuorovaikutuksessa näiden yhä itsenäisempien tekoälyjärjestelmien kanssa, olisi oman etumme mukaista, että ne huomioisivat moraalisia tekijöitä ratkaisuissaan. Tekoälyoptimistit ovat sitä mieltä, että koneista voi tulla joissain suhteissa jopa meitä moraalisesti parempia, koska ne eivät lankea kiusauksiin tai menetä malttiaan (Arkin 2010). Kenties ne oppivat esimerkiksi mallintamaan meitä parhaimmillamme. Tähän tapaan ajattelee yksi Rachel Cuskin romaanissa *Transit* mainituista hahmoista:

> Ystäväni, joka oli masentunut avioeronsa jälkeen, oli hiljattain myöntänyt, että häntä usein liikuttivat kyyneliin asti … automatisoidut äänet junissa ja busseissa, jotka vaikuttivat olevan huolissaan siitä, että hän kulkisi pysäkkinsä ohi. Hän kertoi tuntevansa aidosti jotain rakkauden tapaista sitä naisääntä kohtaan, joka opasti häntä autoa ajaessa niin paljon omistautuneemmin kuin hänen vaimonsa ikinä oli tehnyt. Hän sanoi että olemme keränneet valtavasti kielenkäyttöä ja tietoa elämästä, ja on voinut käydä niin, että vale-inhimillisyydestä on tullut alkuperäistä syvempää ja suhteisiin valmiimpaa, että koneelta voi saada enemmän hellyyttä kuin kanssaihmiseltä. Mekanisoitu käyttöliittymähän tiivistää monta ihmistä, ei vain yhden. (Rachel Cusk, *Transit*, oma käännökseni)

Cuskin kertojahahmon keskustelukumppani on toki tragikoominen hahmo, jota tosikkofilosofin on helppo syyttää projisoinnista ja antropomorfismista. Mutta vaikka emme olisi tosikkoja, monet näkevät keinotekoiseen moraalitoimijuuteen liittyvän periaatteellisia ongelmia. Liityn itsekin tähän kuoroon, mutta koetan myös hieman pohtia, mitä eettisyys vaatii tekoälytoimijoiden valmistajilta ja käyttäjiltä, kun vastuuta ei voi siirtää koneelle. Esitän, että vaikka emme osaisikaan rakentaa aitoon moraaliseen ymmärrykseen kykeneviä

itsenäisiä keinotekoisia toimijoita, meidän on pyrittävä luomaan keinotekoisia *oikeintekijöitä*, järjestelmiä jotka toimivat mahdollisimman pitkälti sillä tavalla kuin täysivaltaiset moraalitoimijat parhaimmillaan.

1. Moraalisesta toimijuudesta

Aloitetaan pohtimalla tarkemmin moraalisen toimijuuden käsitettä ja edellytyksiä. Ensin on huomioitava, että käytämme sanaa ”moraalinen” ja sen sukulaisia sekä kuvailevasti että arvottavasti. Ensimmäisessä mielessä moraalisen toimijan vastakohta ei ole moraaliton toimija, vaan sellainen toimija, joka ei kykene ajattelemaan moraalisesti tai toimimaan sen mukaisesti. Esimerkiksi Eichmann oli kuvailevassa mielessä moraalinen toimija karkeasti siksi, että hänellä oli oma moraalikoodinsa, joka vaikutti hänen tekoihinsa. Arvottavassa mielessä hän ei ollut moraalinen, koska hänen tekonsa olivat hirvittäviä. Korostaakseni tätä eroa käytän yleensä termiä ”moraalitoimija” kun puhun kuvailevassa mielessä. On siis olemassa moraalittomia (eli moraalisesti huonoja) moraalitoimijoita.

Toimijat ylipäänsä erottaa muista olioista ennen kaikkea *tavoitteellisuus*. Toimijat eivät vain vuorovaikuta ympäristönsä kanssa kuten ruostuva rauta, vaan asettavat päämääriä ja tavoittelevat niitä hyödyllisiksi uskomillaan keinoilla. Voimme kenties mielekkäästi sanoa, että sokeria kohti hakeutuvat bakteerit ’toimivat’, mutta tässä tapauksessa *teon* ja *tapahtumisen* rajaviiva on hämärä. Oliot, jotka ovat toimijoita vahvemmassa mielessä, representoivat ympäristöään, päämääriään ja vallassaan olevia keinoja niiden saavuttamiseen, ja nämä representaatiot aiheuttavat ruumiin- tai mielenliikkeitä. Esimerkiksi rotat kykenevät laboratoriokokeiden valossa jopa askelta vaativampaan toimijuuteen, johon kuuluu myös *suunnitelmallisuus*, eli kyky representoida ajallisesti etäisempiä päämääriä ja monivaiheisia keinoja niiden saavuttamiseen, ja pidättäytyä muiden halujen tyydyttämisestä silloin kun se haittaisi päämäärien saavuttamista. Suunnitelmallisuus edellyttää siis jonkinasteista

itsehallintaa ja itsehillintää. Aivohavaintojen mukaan rottienkin on väitetty muun muassa tuntevan katumusta, jos ne sortuvat vähempiarvoisiin pikavoittoihin (Steiner ja Redish 2014).

Mitä suunnitelmalliseen toimijuuteen on lisättävä, jotta saisimme moraalisia toimijoita? Ilmeinen vastaus on, että moraalitoimijan on kyettävä jollain tapaa itsenäisesti arvioimaan mahdollisia päämääriä (eikä vain keinoja) ja vielä tehtävä niin tavalla, joka huomioi moraalin vaatimukset. (Kun puhumme kuvailevassa mielessä, kyse on vaatimuksista, jotka ovat toimijan käsityksen mukaan moraalisia, eikä siten välttämättä aidoista moraalisista vaatimuksista.) Lisäksi tällä arvioinnilla on oltava oikeanlainen merkitys sille, mitä toimija tekee.

Kun lähdemme määrittelemään moraalitoimijuutta tarkemmin, lienee siis hedelmällistä aloittaa seikoista, jotka moraalin näkökulmasta puoltavat tietynlaista toimintaa eli moraalisista perusteista. Esimerkiksi se, että lapsen tönäiseminen kiireessä pois tieltäni tuottaisi vaaratilanteen viattomalle on peruste olla tekemättä sitä. Voimme myös sanoa, että lapselle aiheutuva vaara *antaa* perusteen olla tönäisemättä (vrt. Parfit 2011, 32). Tämä puhetapa on siitä hyödyllinen, että se auttaa erottamaan kaksi tosiasiaa, yhtäältä sen empiirisen faktan, että tietty teko aiheuttaa lapselle vaaraa ja toisaalta sen moraalisen seikan, että vaaran aiheuttaminen lapselle puhuu tekoa vastaan. Käytän ilmausta ”moraalisesti relevantti seikka” niistä tosiasioista, jotka antavat perusteita silloin kun ne vallitsevat, ja termiä ”perustetotuus” siitä korkeamman tason tosiasiasta, että tietty ei-normatiivinen tosiasia antaa perusteen tehdä jokin teko. On täysin mahdollista tunnistaa yksi näistä tunnistamatta toista.

Toinen olennainen erottelu liittyy siihen, kuinka moraalisesti relevantit seikat vaikuttavat toimintaamme. Meidän on mahdollista reagoida havaitsemiimme tai kokemiimme asioihin ilman, että mitenkään pidämme niitä perusteina. Viimeaikainen sosiaalipsykologinen tutkimus on korostanut, että tällaisilla seikoilla saattaa olla yllättävänkin

suuri vaikutus toimintaamme (Doris 2015). On väitetty esimerkiksi, että ihmisen nimi vaikuttaa tilastojen valossa hänen ammatinvalintaansa, tai että tiettyjä asioita tiedostamatta ajattelemaan virittäminen (*priming*) saa esimerkiksi pidättäytymään epärehellisyydestä. On tällä hetkellä empiirisesti kiistanalaista, kuinka merkittäviä nämä vaikutukset ovat, koska useita keskeisiä tuloksia ei ole kyetty toisintamaan. Joka tapauksessa, toinen vaihtoehto on, että tekomme selittää se, että kohtelemme jotain asiaa riittävänä perusteena sille. Joskus, tosin suhteellisen harvoin, teemme näin harkinnan seurauksena. Kenties katson aamulla lämpömittaria ja mietin, minkä takin laittaisin päälle, ja ajattelen, että tällä pakkasella on hyvä syy laittaa ulsteri. Olisi kuitenkin virhe luulla, että perusteena pitäminen vaatii eksplisiittistä harkintaa. Kadulla kiirehtiessäni en missään vaiheessa tietoisesti pysähdy miettimään, tönäisenkö lasta, mutta voin silti kohdella lapselle koituvaa vaaraa perusteena. Tämä näkyy taipumuksissa, jotka toteutuisivat todenvastaisissa tilanteissa (Schlosser 2012, Kauppinen 2015). Jos vahingossa tönäisisin, pyytäisin anteeksi; jos sinä tönäisisit, saattaisin tolvaista sinua. Lapselle aiheutuva haitta ei ole vain syy sille, miksi en tönäise, vaan myös minun itseni hyvänä pitämä syy sille.

Tarkoituksemme takia on olennaista erottaa nimenomaan *moraalisena* perusteena kohteleminen perusteena kohtelemisesta ylipäänsä. Nähdäkseni moraaliset perusteet liittyvät niin sanottuihin *reaktiivisiin asenteisiin*, kuten suuttumukseen, halveksuntaan, kiitollisuuteen, syyllisyyteen, häpeään ja ylpeyteen. Kuten esimerkiksi Edward Westermarck (1906), Peter Strawson (1962) ja Stephen Darwall (2006) ovat korostaneet, moraalisen ajattelun keskeinen funktio on nimenomaan tällaisten asenteiden ohjaaminen ja ilmaiseminen, ja sitä kautta toimintaan vaikuttaminen. Jos esimerkiksi näen jonkun toimivan tavalla, jota vastaan on mielestäni riittävät moraaliset perusteet, olen taipuvainen kielteisiin reaktiivisiin asenteisiin tai ainakin niiden pitämiseen sopivina. Minun näkökulmastani tekijän on paikallaan hävetä

ja muiden sopii suuttua. Tämä erottaa perusteiden kohtelemisen moraalisina niiden kohtelemisesta esimerkiksi prudentiaalisina (omaan etuun liittyvinä), esteettisinä tai laillisina.

On houkuttelevaa ajatella, että ollaksemme moraalitoimijoita, meidän on kyettävä tunnistamaan moraaliset perusteet ja huomioimaan ne oikein toiminnassamme. Mutta tämä vaatii selvennystä edellisten erottelujen valossa. Niiden havainnollistamiseksi voimme vertailla eri versioita tilanteesta, jossa lapsi kävelee hitaasti kiireisen aikuisen edessä vilkkaasti liikennöidyn tien vieressä:

- Ailo tönäisee lapsen kadulle, koska hän kärsii harhoista ja luulee lasta demoniksi, ja hänellä on pakkomielle päästä demoneista eroon.
- Ben tönäisee lapsen kadulle, koska hän luulee virheellisesti, että se on lapsesta kivaa eikä ole vaarallista.
- Cersei tönäisee lapsen kadulle, koska hänestä meillä on riittävän hyvä syy satuttaa lapsia.
- Desi jättää tönäisemättä lapsen kadulle, koska lapsen takki on luminen, eikä hän halua kastella kalliita nahkahanskojaan.
- Elina jättää tönäisemättä lapsen kadulle, koska hänellä on tiedostamaton taipumus vältellä toisten satuttamista.
- Frank jättää tönäisemättä lapsen kadulle, koska se aiheuttaisi lapselle vaaratilanteen.

Oletetaan, että tässä tilanteessa lapsen tönäiseminen on moraalisesti väärin. Silloin Ailo ja Ben toimivat moraalisesti väärin, koska heillä on virheellinen uskomus moraalisesti relevantista seikasta. Cersei toimii väärin, koska hänellä on virheellinen uskomus perustetotuuksista. Desi toimii oikein, mutta vain kausaalisesti vaikuttavan tekijän takia, ja Elina väärästä syystä. Vain Frank tekee oikein oikeasta syystä.

Kutsun ensimmäistä kolmea toimijaa *väärintekijöiksi* ja jälkimmäisiä kolmea *oikeintekijöiksi*. Nämä ovat kattotermejä niille henkilöille, jotka toimivat väärin tai oikein, riippumatta siitä miksi he tekevät niin. Jos Ailo säännöllisesti ja ilman omaa syytään kärsii vakavista harhoista tai pakkomielteistä, hän ei kykene vastaamaan moraalisiin perusteisiin, eikä häntä voi pitää moraalisesti vastuullisena toimijana. Ben ja Cersei sen sijaan ovat todennäköisesti kykeneviä toimimaan oikein, jos he vaan käyttävät asianmukaisesti kykyjään muodostaa empiirisiä tai moraalisia uskomuksia. He ovat siten moraalisesti vastuussa virheestään, ellei sille ole jotain anteeksiantoperustetta (kuten omasta laiminlyönnistä johtumatonta tietämättömyyttä tai ulkoista pakotusta).

Aiheemme näkökulmasta meitä kiinnostavat erityisesti erityyppiset *oikeintekijät*. (Jostain syystä tämä sana puuttuu sekä suomen kielestä että monista muista, kuten englannista.) Voi olla, että tavalliset moraalitoimijat toimivat joskus moraalisesti hyväksyttävällä tavalla syistä, joita he eivät tiedosta. Mutta entä jos Desi säännöllisesti tekee oikein vain siksi, että tilannetekijät, joita hän ei tunnista, vain sattuvat vaikuttamaan häneen siten? On nähdäkseni täysin mahdollista, että olisi oikeintekijöitä, jotka eivät kykene moraaliseen ajatteluun. Voimme silloin sanoa, että Desi vastaa aitoon moraaliseen perusteeseen *de re*, mutta ei *de dicto* – kohtelematta sitä nimenomaan moraalisena perusteena. Jos hän ei kykene muuhun, hän ei ole moraalitoimija, vaikka olisi hyvällä tuurilla oikeintekijä. Mikäli taas Elina toimii säännöllisesti tässä kuvatulla tavalla, eli tekee oikein siksi, että se sattuu hänen näkökulmastaan lankeamaan yksiin hänen oman etunsa kanssa, hän on todennäköisemmin moraalitoimija, joka tekee oikein väärästä syystä. Näin sanoessamme oletamme, että hän kykenisi tekemään toisinkin ja vastaamaan perustetotuuksiin. Jos hän sen sijaan on aidosti sokea muille kuin omaa etuaan koskeville perustetotuuksille, kuten jotkut psykopaatit kenties ovat, häntä ei voi pitää moraalisesti täysin vastuullisena. Kuten Kant korosti, on merkittävä ero sen välillä, toimimmeko moraalin vaatimusten *mukaisesti* vai niiden *ohjaamina*.

Moraalitoimijat vähintään kykenevät jälkimmäiseen, kun pelkät oikeintekijät pääsevät vain edelliseen. Viimeisenä, jos Frank tekee säännöllisesti oikein oikeasta syystä, hän on tietysti esimerkki hyvästä moraalisesta toimijasta.

Mitä moraalitoimijuus siis edellisen valossa vaatii? Selvästikin moraalitoimijan on kyettävä kohtelemaan joitakin seikkoja moraalisina perusteina ja toimimaan niiden mukaisesti, ja jossain määrin myös tehtävä näin. Toisella tavalla ja kenties hieman harhaanjohtavastikin ilmaistuna, hänen on kyettävä erottamaan moraaliset säännöt ei-moraalisista ja noudattamaan niitä. Mutta mikä sitten on leimallista moraalisille perusteille ja säännöille? Ottaen huomioon, että kysymyksemme ei koske *aitoja* perusteita tai *oikeita* sääntöjä, olisi virhe painottaa vastauksessa tiettyjä sisältöjä, kuten toisten edun huomioimista. Olennaista on sen sijaan, että moraali liittyy *reaktiivisiin asenteisiin*, kuten jo totesin. Moraalitoimija kohtelee joitakin tekoja ja luonteenpiirteitä sen mukaisesti, että ne ansaitsevat sellaisia asenteita kuin paheksunta tai halveksunta, mikäli tekijä on vastuussa niistä. Tämä ilmenee muun muassa siinä, että hän jättää teon tekemättä, vaikka siitä olisi hänelle hyötyä, tai syyllisyyden tuntemisena. Sanon lyhyesti, että jos olen moraalitoimija, sillä teenkö oikein vai väärin on *väliä* minulle.

Perusteiden ja sääntöjen kohtelemiseen moraalisina liittyy myös ajatus niiden yleispätevyydestä – siitä että niiden auktoriteetti ei perustu jonkun valtaan tai oman etuni edistämiseen. Kuten Hume laittamattomasti sanoi:

> Kun joku nimeää toisen *vihollisekseen*, *kilpailijakseen*, *vastustajakseen* tai *vastapuolekseen*, hänen ymmärretään puhuvan itserakkauden kieltä ja ilmaisevan tunteita, jotka eroavat muiden vastaavista ja juurtavat hänen erityisistä olosuhteistaan ja tilanteestaan. Mutta kun hän luonnehtii jotakuta *paheelliseksi* tai *katalaksi* tai *turmeltuneeksi*, hän puhuu toista kieltä ja ilmaisee tunteita, jotka hän odottaa yleisönsä tulevan jakamaan. Hänen täytyy siten nousta yksityisen ja erityisen tilanteensa

> yläpuolelle ja valita näkökulma, joka on yhteinen hänelle ja toisille. (Hume 1751, oma käännökseni)

Hume luonnehtii tässä asiaa sentimentalistisessa sävellajissa, mutta yleisemmin muotoiltuna ajatuksen hyväksyvät monen eri metaeettisen koulukunnan edustajat. Moraalitoimijuuteen ei tietenkään riitä, että kykenemme moraalisiin arvostelmiin. Ne on myös kyettävä panemaan toimeen. Se edellyttää suunnitelmallisuutta ja itsehallintaa, mikä korostuu siksi, että moraaliseksi koetut säännöt usein vaativat luopumaan jostain, mitä muuten haluaisimme.

Moraalitoimijuudessa on luontevaa erottaa eri asteita. On *minimaalisia* moraalitoimijoita, jotka omaksuvat periaatteensa ja asenteensa toisilta kykenemättä kyseenalaistamaan oppimaansa. He ajattelevat kyllä moraalisesti ja enemmän tai vähemmän toimivat käsitystensä mukaisesti, mutta joutuvat ymmälle, jos normeilla on ristiriitaisia implikaatioita, jos niitä pitää soveltaa luovasti, tai jos joku haastaa puolustamaan niitä. Esimerkiksi lapset ja joidenkin väitteiden mukaan jotkut autistit kuuluvat tähän ryhmään (ks. Kauppinen 2017a). Minimaalisten toimijoiden moraalinen ja episteeminen kompetenssi on vaillinainen. Sen takia he eivät välttämättä ole täydessä vastuussa tekemisistään – jos he tekevät väärin, kyse voi olla aidosti kyvyttömyydestä tehdä oikein, vaikka oikein tekemisellä olisi heille riittävästi väliä.

Sen sijaan *täysivaltaiset* moraalitoimijat kykenevät *ymmärtämään,* miksi jotkut teot ovat väärin, ja tämä ymmärrys voi myös motivoida heitä asianmukaisesti. He kykenevät muodostamaan itsenäisiä arvostelmia, eivätkä siten ole riippuvaisia toisista. Ymmärtämiseen relevantissa mielessä sisältyy muun muassa osien roolin kokonaisuudessa ja niiden välisten riippuvuussuhteiden hahmottaminen, mitkä mahdollistavat vastaamisen kysymyksiin siitä, minkä pitäisi muuttua teon moraalisen statuksen muuttamiseksi (Hills 2009, Grimm (toim.) 2017). Otetaan esimerkiksi, että nyrkkisääntönä on väärin kertoa rasistisia vitsejä. Tällä

lienee tekemistä sen kanssa, että se ilmentää ansaitsematonta ylenkatsetta, ulkopuolistaa kohteita ja usein vahvistaa heihin kohdistuvia ennakkoluuloja. Siten siinä on kyse jo sinänsä eriarvoisesta kohtelusta, joka antaa aihetta loukkaantua. Mutta mitä vikaa on eriarvoisessa kohtelussa tai ylenkatseessa tai loukkaavassa käytöksessä ja miksi? Miksi nämä seikat puhuvat tekoa vastaan? Milloin kenties on hyväksyttävää kertoa rasistisia vitsejä? Jos ymmärtää, miksi teko on väärin, osaa periaatteessa vastata tällaisiin kysymyksiin, koska on saanut tiukan otteen moraalitotuuksista – on *oivaltanut*, miksi ne ovat niin kuin ovat.

Voimme erottaa karkeasti kolme filosofista koulukuntaa sen suhteen, mitä moraalinen ymmärrys vaatii. *Intuitionistit* ajattelevat, että kykenemme ymmärryksellä välittömästi tavoittamaan moraalisia totuuksia, joko yleisiä tai erityisiä. Esimerkiksi Rossin (1930) ja Audin (2004) mukaan on itsestään selvää, että lupaukset on lähtökohtaisesti pidettävä ja vahingon tekemistä vältettävä. Kuka tahansa, joka riittävän hyvin ymmärtää nämä moraaliset väitteet, on myös oikeutettu uskomaan niiden totuuteen. Ne ovat synteettisiä a priori totuuksia, joita ei voi johtaa mistään perustavammasta. Niiden ymmärrystä voi kyllä edistää pohtimalla erityistapauksia, joissa ne ilmenevät. *Rationalistit* puolestaan uskovat, että voimme vain järkeä käyttämällä saada tukevan otteen moraalisista totuuksista. Kaikkein kuuluisimmin Kant ajatteli karkeasti niin, että järki vaatii meitä toimimaan yleisiksi laeiksi soveltuvien toimintaperiaatteiden mukaisesti, ja että voimme a priori todeta, millaiset periaatteet johtaisivat yleistettyinä väistämättä jonkinlaiseen ristiriitaan, mikä tekee niistä järjenvastaisia. Viimeisinä, muttei suinkaan vähäisimpinä, *sentimentalistit* uskovat, että moraalinen ymmärrys vaatii sopivia tunnereaktioita, erityisesti empatiaa toisten kärsimystä tai reaktiivisia asenteita kohtaan, ja näiden tunnereaktioiden hallintaa erilaisten vinoumien välttämiseksi. Sentimentalistien on helppo selittää, miksi jonkun teon ymmärtäminen vääräksi motivoi olemaan tekemättä sitä, koska vältämme yleensäkin kielteisiä tunteita herättäviä tekoja.

Sentimentalismin viehätystä voi ehkä ymmärtää, jos ajattelee mitä vaikkapa tuskan ymmärtäminen itsessään huonoksi asiaksi oikeastaan vaatii. Otetaan Frank Jacksonin (1981) kuuluisaa ajatuskoetta mukaillen esimerkiksi Mary, joka ei ole koskaan kokenut nälkää eikä mitään muutakaan epämiellyttävää tilaa. Mary voi kyllä kuulopuheiden perusteella ymmärtää, että nälällä on huonoja seurauksia, kuten keskittymisen ja monenlaisen toiminnan haittaaminen. Hänen on myös mahdollista *tietää*, että se on itsessään huono asia, jos joku luotettava taho kertoo niin (Enoch 2014). Mutta kun Mary ensimmäistä kertaa itse näkee nälkää, hän vaikuttaisi oppivan jotain, eikä vain siitä, miltä nälkä tuntuu, vaan myös jotain sen huonoudesta. ”*Nyt* tajuan, miksi tämä on niin paha juttu!”, hän voi sanoa. Samaan tapaan joku voi tietää, että kärsimyksen aiheuttaminen ilman syytä on väärin, mutta ymmärtää tai oivaltaa *miksi* vasta kun asettuu uhrin asemaan.

Vaikka olenkin yleisesti ottaen sentimentalisti (Kauppinen 2017b), en tässä oleta minkään näistä malleista totuutta. Mutta jotta moraalitoimijuuden merkitys vielä selkiytyisi, vertaan vielä tämän osuuden lopuksi kahta erilaista oikeintekijää, joista toinen ei ole lainkaan moraalitoimija ja toinen on täysivaltainen sellainen. Ensimmäinen voi olla vaikkapa psykopaatti, joka kirjaimellisesti ei kykene ymmärtämään moraalin pointtia psykologisten rajoitteidensa takia. Tällainen henkilö voi kuitenkin periaatteessa oppia, mitkä sosiaalisesti opetetut säännöt ovat moraalisia – mitä yleensä paheksutaan. Hän voi myös tulla siihen tulokseen, että on hänen oman etunsa mukaista noudattaa moraalisia sääntöjä, ja luettuaan Hobbesia hän voi haluta kaikkien muidenkin noudattavan niitä, koska siitä on ei-moraalisia hyötyjä, joten hän on valmis rankaisemaan väärintekijöitä. Jos vielä oletamme, että hänelle satutaan opettamaan juuri oikeat säännöt, hän saattaa olla virheetön oikeintekijä. Selvästi hän kuitenkin eroaa esimerkiksi sellaisesta toimijasta, joka hyväksyy samat säännöt siksi, että ymmärtää niiden jujun esimerkiksi hallitun empatian kautta. Tämä ero näkyy edellä mainittujen sovellus- ja perustelutaipumusten lisäksi myös oikeintekemisen *takuuvarmuudessa*.

Oikeintekevä täysivaltainen moraalitoimija tekisi oikein myös sellaisissa todenvastaisissa tilanteissa, joissa hänet olisi opetettu tekemään väärin, joissa toiset tosiasiassa tekisivät väärin ja joissa hänen yleinen motivaatiotilansa olisi erilainen, koska hänen motivaationsa oikein tekemiseen juontaa ymmärryksestä itsestään. On siis monta syytä, miksi ei ole yhdentekevää, olemmeko täysivaltaisia moraalitoimijoita, vaikka kaikissa tilanteissa ero ei näy tosiasiallisessa toiminnassa.

2. Keinotekoiset moraaliset toimijat?

Nyt kun meillä on selvempi käsitys moraalisen toimijuuden edellytyksistä, voimme siirtyä tarkastelemaan keinotekoisen moraalitoimijuuden mahdollisuutta. Peruskuvio on, että pessimistit argumentoivat, että koneelta puuttuu joko vääjäämättä tai näköpiirissä olevassa tulevaisuudessa jokin moraalitoimijuuden välttämätön edellytys. Optimistit taas esittävät evidenssiä tätä vastaan tai argumentoivat, ettei kyseinen piirre itse asiassa ole olennainen moraalitoimijuudelle relevantissa mielessä.

Jätän tässä syrjään yleisemmät huolenaiheet siitä, kykenevätkö koneet suuntautumaan maailmaan ja aidosti tavoittelemaan päämääriä (kuten Searlen 1981 kritiikin). Vaikka filosofien esittämät huolenaiheet ovat varteenotettavia, on joka tapauksessa mielenkiintoista pohtia, liittyykö keinotekoisen moraaliseen toimijuuteen erityisiä ongelmia. Aloitan täyttä moraalitoimijuutta vastaan esitetyistä argumenteista. Ensimmäinen vetoaa vapaaseen tahtoon:

1. Jos S on täysivaltainen moraalitoimija, S on moraalisesti vastuussa teoistaan.
2. Moraalinen vastuu teoista edellyttää vapaata tahtoa.
3. Keinotekoisilla toimijoilla ei ole vapaata tahtoa.
4. Siis, keinotekoiset toimijat eivät ole moraalisesti vastuussa teoistaan.
5. Siis, keinotekoiset toimijat eivät ole täysivaltaisia moraalitoimijoita. (vrt. Himma

2008)

En käytä tämän argumentin tarkasteluun paljoa aikaa. Moraalitoimijuuden yhteys vapaaseen tahtoon kulkee vastuun kautta, joten jätän tässä syrjään syvät kysymykset vapaan tahdon mahdollisuudesta ja todellisuudesta (ks. Visala 2018) ja keskityn vastuun itsensä edellytyksiin. Yksi perustelu tälle on se, ettei ole lainkaan harvinaista kiistää vastuun edellyttävän vapaata tahtoa (ja siten hylätä premissi 2). Näin tekevät muun muassa John Martin Fischer ja Mark Ravizza (1998), jotka ovat valmiita hyväksymään, ettei meillä ole vapaata tahtoa ainakaan kovin vahvassa mielessä, mutta argumentoivat, että voimme siitä huolimatta olla moraalisesti vastuullisia silloin kun tekomme voidaan lukea meille itsellemme.

Toinen argumentti on sukua edelliselle, mutta metafyysisesti vaatimattomampi:

1. Jos S on täysivaltainen moraalitoimija, S on moraalisesti vastuussa teoistaan.
2. Moraalinen vastuu teoista edellyttää autonomian kanssa yhteensopivaa historiaa.
3. Keinotekoisilla toimijoilla ei ole autonomian kanssa yhteensopivaa historiaa, koska ne on esiohjelmoitu toimimaan/oppimaan tietyllä tavalla.
4. Siis, keinotekoiset toimijat eivät ole moraalisesti vastuussa teoistaan.
5. Siis, keinotekoiset toimijat eivät ole täysivaltaisia moraalitoimijoita.

Tässä lähtökohtana on se, että autonomiaan ei riitä esimerkiksi se, että haluamme tehdä niitä asioita, joita arvostamme tai joita haluamme haluta tehdä. Onhan mahdollista, että arvomme tai korkeamman asteen halumme ovat seurausta jonkinlaisesta tietoisesta manipulaatiosta. Patrick Hew (2014) ja Raul Hakli ja Pekka Mäkelä (2016) ovat hieman eri tavoin esittäneet, että esiohjelmointi vastaa moraalisesti tällaista manipulaatiota. Kuten Hakli ja Mäkelä asian ilmaisevat, “robotit eivät voi olla moraalisesti vastuullisia, koska toiset toimijat ovat suunnitelleet ja ohjelmoineet niille sellaisen ‘luonteen’ kuin niillä on”. Al Mele (1995), jonka työhön he nojaavat, korostaa ihmisten tapauksessa, että tässä on olennaista,

että manipuloidut tai esiohjelmoidut arvot ovat "käytännöllisesti lukkoonlyötyjä" – toimijat eivät kykene rationaalisesti ja vapaaehtoisesti muuttamaan niitä, koska tämä edellyttäisi jonkinlaista pääsyä niiden ulkopuolelle. Tämä on lupaava argumentti, mutta en kuitenkaan pidä sitä yksin ratkaisevana. On ensinnäkin oltava tarkkoja siitä, ettei rima nouse niin korkealle, etteivät useimmat aikuiset ihmisetkään ylitä sitä. Toiseksi, ainakin Hakli ja Mäkelä ovat valmiita olettamaan, että synkroniset moraalisesti relevantit kyvyt voitaisiin toteuttaa keinotekoisesti. Jos näin pystyttäisiin tekemään, en näe mitään syytä, miksi keinotekoiset toimijat eivät voisi päästä eroon esiohjelmoiduista arvostuksista aivan samassa määrin kuin luonnolliset moraalitoimijatkin, jotka kykenevät esimerkiksi muuttamaan asenteitaan erirotuisia ihmisiä kohtaan henkilökohtaisen kanssakäymisen perusteella. Keinotekoiselle toimijalle annettu sysäys oikeaan suuntaan on nähdäkseni yhteensopiva moraalisen vastuun kanssa siinä kuin lapsen moraalinen opettaminenkin. Mikäli lapsi jossain vaiheessa tavalla tai toisella kypsyy itsenäiseen moraaliseen ymmärrykseen, hän ei enää ole vain hänelle opetettuja näkemyksiä toisteleva papukaija.

Mutta onko moraalisten kykyjen keinotekoinen toteuttaminen realistinen tavoite? Tässä yksi argumentti sitä vastaan:

1. Täysivaltainen moraalitoimijuus vaatii moraalista ymmärrystä.
2. Moraalinen ymmärrys edellyttää hallittuja tunteita/arvostelukykyä.
3. Keinotekoisilla toimijoilla ei ole hallittuja tunteita/arvostelukykyä.
4. Siis, keinotekoiset toimijat eivät ole täysivaltaisia moraalitoimijoita.

Ensimmäisen ja toisen premissin puolesta esitin jo aiemmin joitakin seikkoja, jättäen avoimeksi mikä on paras malli moraalisesta ymmärryksestä tai kompetenssista. Sentimentalismin mukaan ymmärtääkseen moraalin jujun on karkeasti sanoen kyettävä empatiaan ja tunteiden hallintaan laajemmasta perspektiivistä (Kauppinen 2017b). Lienee

varsin uskottavaa, että tällaisen ymmärryksen keinotekoisesta tuottamisesta ollaan kaukana. Yksi syy olla skeptinen sen suhteen on, ettei keinotekoisten tunteiden tuottamisessa olla edistytty lainkaan samaan aikaan kun keinoäly on muuten ottanut valtavia edistysaskeleita. (Palaan tähän kohta.) Joku voisi ajatella, että samasta syystä rationalistinen malli antaa syytä toiveikkuuteen. Mutta olen skeptinen tämänkin suhteen. Kuten kaikki Kantin tai kantilaisten argumentteihin perehtyneet tietävät, kyse ei ole mistään suoraviivaisesta loogisesta päättelystä, vaan maksiimien yleispätevyyden koetteleminen vaatii kokonaisvaltaista arvostelukykyä ja jättää sijaa tulkinnalle ja kiistelylle. Esimerkiksi Derek Parfit (2011) vetoaa vaikutusvaltaisessa kantilaisen näkemyksen uudelleenmuotoilussaan jatkuvasti arvostelmiin perusteista ja niiden vahvuudesta yrittämättäkään johtaa niitä muodollisista ehdoista. Tämä tuo rationalismin lähemmäksi intuitionismia, joka ei myöskään tarjoaa mitään *menetelmää* itsestään selvien synteettisten totuuksien tunnistamiseksi, vaan vetoaa jonkinlaiseen arvostelukykyyn tai ”riittävään ymmärrykseen” (Audi 2004) ja joskus hyveelliseen ”näkemiseen”, jolle ei voi antaa kuvailevia kriteerejä (esim. McDowell 1979). Tätä tietynlaista epämääräisyyttä voi toki pitää näiden näkemysten filosofisena ongelmana, mutta ainakin se viittaa siihen, että moraalinen kompetenssi vaatii juuri sellaista laajakatseista arvostelukykyä, jonka toteuttaminen vähemmänkin vaativien kysymysten suhteen on osoittautunut erityisen vaikeaksi, vaikka affektiivisuuden haasteet jätettäisiin syrjään.

Olen tähän mennessä puhunut täysivaltaisen moraalitoimijuuden toteuttamisen haasteista. Mutta entä minimaalisemmin ymmärretty toimijuus, jonka tärkeimpänä edellytyksenä on ylipäänsä joihinkin asioihin suhtautuminen leimallisesti moraalisella tavalla ja sen mukaisesti toimiminen? Keskityn tässä vain yhteen argumenttiin:

1. Moraalitoimijuus vaatii kykyä kohdella joitakin perusteita moraalisina.
2. Perusteiden kohteleminen moraalisina edellyttää, että välittää siitä, tekeekö oikein vai väärin.

3. Välittäminen edellyttää sitä, että tuntee jotain (kokemus- eli fenomenaalista tietoisuutta).
4. Keinotekoisilta toimijoilta puuttuu kokemustietoisuus.
5. Siis, keinotekoiset toimijat eivät voi olla moraalitoimijoita.

Olen puolustanut aiemmin kahta ensimmäistä premissiä, ja moni kenties hyväksyy neljännen. Mutta entä kolmas? Jotkut ovat valmiita kiistämään sen. Wendell Wallach ja Colin Allen (2008) esittävät, että funktionaalinen samankaltaisuus riittää:

> Jotkut filosofit pitävät kiinni siitä, että *fenomenaalinen* tietoisuus vaatii jotain funktionaalisen samankaltaisuuden ylittävää, eikä tietokoneiden onnistuminen inhimilliseen tietoisuuteen liittyvien tehtävien suorittamisessa tule ikinä tyydyttämään heitä. Mutta tämä käsitys tietoisuudesta asiana, joka ei mitenkään vaikuta havaittavaan käyttäytymiseen, on irrelevantti keinotekoisten moraalisten toimijoiden kehittämisen kannalta. *Vain funktionaalisella samankaltaisuudella voi olla väliä keinotekoisten moraalisten toimijoiden suunnittelemiselle.*

Ydinajatuksena tässä on nähdäkseni se, että moraalitoimijuudelle riittää, jos robotit käyttäytyvät *ikään kuin* olisivat tietoisia, tai että ne ovat tietoisia jossain muussa mielessä kuin kokemuksellisessa. Jotkut tieteilijät uskovat, että tietoisuuden luominen on jo näköpiirissä. Owen Holland on esittänyt, että keinotietoisuutta voi pyrkiä rakentamaan kolmella eri menetelmällä: tunnistamalla ja mallintamalla tietoisuuden osatekijät, mallintamalla tietoisten olentojen aivot tai luomalla olosuhteet, joissa tietoisuus kehittyy itsestään (Wallach ja Allen 2008). Tutkimus on paljolti keskittynyt ensimmäiseen näistä. Esimerkiksi Stan Franklinin LIDA (*learning intelligent distribution agent*) on malli, joka lähtee Bernard Baarsin (1997) ajatuksesta, että tietoisuus on eräänlainen ”globaali työtila”, jonka hallinnasta

tiedostamattomien prosessien yhteenliittymät kilpailevat. Franklinin lähinnä käsitteellisessä mallissa ohjelmalliset prosessit, jotka vastaavat erilaisia ulkoisia ja sisäisiä havaintosyötteitä ja erityyppisiä muisteja välittävät informaatiota huomiota mallintaville alarutiineille, jotka puolestaan nostavat korkeimmalle rankatun informaation yleiseen työtilaan, josta se välittyy kaikille aliohjelmille. Franklin (2003) esittää, että tietoisuuden funktio on esimerkiksi auttaa toimimaan uusissa tilanteissa, varoittaa vaaroista, kertoa toimintamahdollisuuksista ja mahdollistaa ympäristön piirteisiin sopiva käyttäytyminen, ja uskoo että hänen mallinsa toteuttava järjestelmä toimii vastaavasti.

Tämä herättää paljon kysymyksiä, jotka liittyvät niin sanottuun tietoisuuden vaikeaan ongelmaan, eli sen selittämiseen, kuinka fysikaaliset prosessit voivat aiheuttaa tai toteuttaa asioiden subjektiivisen tunnun (Chalmers 1996, Pylkkänen 2007). Tässä yksi pikainen argumentti kokemustietoisuuden irrelevanssia vastaan:

1. Jos on mahdollista jäänteettä jäljitellä tietoisuutta funktionaalisesti ilman fenomenaalista tietoisuutta, fenomenaalinen tietoisuus on *epifenomenaalista*. (Toisin sanottuna, zombielta ei puutu mitään, mikä vaikuttaisi hänen toimintaansa.)
2. On hyvin epätodennäköistä, että fenomenaalinen tietoisuus on epifenomenaalista.
3. Siis, on hyvin epätodennäköistä, että tietoisuutta voi jäänteettä jäljitellä funktionaalisesti ilman fenomenaalista tietoisuutta.

Yksi syy uskoa toiseen premissiin juontaa evoluutiosta. On periaatteessa mahdollista, että kokemustietoisuus on luonnonvalinnan sivutuote. Mutta sellaiseksi se vaikuttaa vaativan varsin monimutkaista arkkitehtuuria. On paljon todennäköisempää, että siitä on adaptiivista hyötyä meille ja muille eläimille, mikä taas edellyttää, että sillä on toiminnallinen rooli eli että se ei ole epifenomenaalista. Lyhyesti, olisi melkoinen ihme, että meillä on kokemustietoisuus, jos sillä ei ole jotain tärkeää funktionaalista roolia. Kenties se liittyy siihen, jota joskus

kutsutaan alkuperäiseksi intentionaalisuudeksi. Tästä seuraa tietysti, ettei ole mahdollista, että zombiet käyttäytyisivät aina täysin samoin kuin vastaavassa tilanteessa olevat kokemustietoiset yksilöt. Tieteellistä näyttöä tästä ei ole, mutta elokuva- ja televisiozombiet vaikuttavat kyllä varsin kylmäkiskoisilta!

Tähän voi toki vastata siten, että saman funktion voi yleisesti ottaen toteuttaa eri tavoilla. Kuten Wallach ja Allen (2008) huomauttavat, Deep Blue voitti Kasparovin pelaten aivan eri periaatteilla, joten miksei tunteiden ja kokemustietoisuuden moraalista funktiota voisi korvata toisella tavalla toimivalla järjestelmällä? Tämä olisi kenties mahdollista, jos niillä olisi vain välineellinen rooli. Mutta siinä määrin kuin ne ovat *olennaisesti konstitutiivisia* välittämiselle ja asioiden moraalisena kohtelemiselle, niitä ei voi korvata jollakin muulla arkkitehtuurilla.

Tämä jättää avoimeksi sen vaihtoehdon, että luomme keinotekoiselle toimijalle keinotekoiset tunteet. Mutta tämäkin on erittäin haastava tehtävä. Vaikka on olemassa esimerkiksi sellainen tieteenala kuin affektiivinen tietojenkäsittelytiede, sen tärkeimpiä päämääriä ovat tekoälyn opettaminen tunnistamaan tunteita ja vastaamaan niihin soveliaasti, ei synteettisten tunteiden luominen sinänsä (Picard 1995). Monet ovat skeptisiä jälkimmäisten mahdollisuuden suhteen. Esimerkiksi Steve Torrance (2008) ja Amanda Sharkey (2017) argumentoivat, että tunteet ja ylipäänsä kokemuksellisuus ovat olennaisesti ruumiillisten, itseorganisoituvien biologisten organismien ominaisuuksia.

Vastauksena tämäntyyppisiin argumentteihin, jotkut tekoälyoptimistit luopuvat suosiolla ihmisenkaltaisen moraalisen toimijuuden jäljittelystä ja argumentoivat sen sijaan, että koneet voisivat olla moraalisia toimijoita jossain laajemmassa mielessä – eräänlaisia ersatz-moraalitoimijoita. Esimerkiksi Luciano Floridi ja Jeff Sanders (2004) lähtevät siitä, että voimme tarkastella järjestelmiä eri abstraktiotasoilla, ja abstraktiotason valinta vaikuttaa siihen, ovatko ne toimijoita vai eivät. Sopivalla abstraktiotasolla myös tekoälyjärjestelmät

täyttävät heidän ehtonsa toimijuudelle, koska ne vuorovaikuttavat ympäristönsä kanssa sisäisten tilojensa ohjaamina ja voivat muuttaa reaktioitaan ohjaavia sääntöjä. Mikäli niiden teot aiheuttavat ”moraalista hyvää tai pahaa”, ne ovat Floridin ja Sandersin mukaan moraalitoimijoita. He myöntävät, ettei koneita voi kuitenkaan pitää moraalisesti vastuullisina, koska niitä ei ole mielekästä rangaista. Ne voivat kuitenkin olla ”tilivelvollisia” (*accountable*) esimerkiksi siinä mielessä, että jos ne toimivat väärin, ne kannattaa korjata tai purkaa. Mark Coeckelbergh (2014) ja David Gunkel (2018) puolestaan lähtevät siitä, että meidän tulisi lähestyä moraalitoimijuutta ja -subjektiutta relationaalisesti, lähtien siitä, mitä asioita kohtelemme moraalisina, riippumatta niiden sisäsyntyisistä määreistä. Coeckelbergh tiivistää, että “Tässä lähestymistavassa ei enää ole kuilua ‘oikean’ tavan nähdä robotti ja minun ‘havaintoni’ robotista välillä.” (2014, 71)

Nämä yritykset ovat monella tapaa ongelmallisia. Kummatkin relativisoivat toimijuuden joko intresseihimme tai suhtautumistapoihimme. On toki totta, että voi olla joitakin tarkoituksia varten hyödyllistä puhua erilaisista järjestelmistä toimijuuden kielellä, kuten Dennett (1987) jo kauan sitten huomautti. Mutta aiheuttaa vain sotkua, jos hämärrämme eron itsenäisesti päämääriä asettavien ja tiettyjen parametrien rajoissa päämääriään palautteen perusteella säätävien olioiden välillä. Olen jo argumentoinut, ettei jälkimmäisistä tee moraalitoimijoita niiden tekojen seurausten moraalinen relevanssi. Moraalitoimijuuteen kuuluvat ainakin vastuun kannalta välttämättömät kyvyt. Se mistä Floridi ja Sanders puhuvat ei tosiasiassa ole missään mielessä moraalinen tilivelvollisuus, vaan ainoastaan kausaalinen vastuu. On olennaista esimerkiksi huollon takia kyetä tunnistamaan, mitä muuttaa, jos järjestelmä toimii epätoivottavasti, mutta puhe tilivelvollisuudesta vain sumentaa asiaa. Cockelberghin ja Gunkelin kriteerit taas ovat niin väljiä, että niiden mukaan lasteni pehmolelut olisivat ilmeisesti moraalisia subjekteja. Jos he olisivat oikeassa, olisi käsitteellisesti mahdotonta erehtyä siitä, onko joku moraalitoimija, jos

sitä sellaisena pidetään. Mutta näin ei suinkaan ole. On joitakin asioita, joiden suhteen olemme ainakin kollektiivisesti periaatteellisesti erehtymättömiä (emme voisi kaikki olla väärässä siitä, mikä on muodikasta), mutta ei ole hyvää syytä ajatella, että moraalitoimijuus lukeutuisi näihin.

3. Kohti keinotekoisia oikeintekijöitä

Olen esittänyt, että keinotekoisten moraalitoimijoiden luomiselle on monia haasteita. Mutta sanomani ei ole pelkästään negatiivinen. Vaikka hyvä moraalitoimija on takuuvarmempi kuin pelkkä oikeintekijä, kuten aiemmin argumentoin, olisimme jo varsin pitkällä jos pystyisimme luomaan tavallisissa oloissa luotettavia keinotekoisia oikeintekijöitä. Tärkeintä on kuitenkin, että koneet toimivat moraalisesti oikealla tavalla, tai ainakin tavalla, joka olisi oikea, jos moraalitoimija tekisi teon samassa tilanteessa. (Tätä viittausta moraalitoimijan kaltaiseen toimintaan tarvitaan, koska ei ole kiistatonta, voimmeko ylipäänsä mielekkäästi sanoa, että jonkun, joka ei ole moraalitoimija, *pitää* tehdä mitään, koska monet moraaliteoriat liittävät pitämisen tilivelvollisuuteen anteeksiantoperusteen puuttuessa (esim. Darwall 2006). Ei ole selvää pitääkö tämä paikkansa – vaikuttaisihan täysin mielekkäältä sanoa, että psykopaatti, joka ei kykene erottamaan oikeaa väärästä eikä siten ole vastuullinen, voi kuitenkin tehdä oikein tai väärin.)

> Keskeinen normatiivinen väitteeni on seuraava vastuuperiaate:
>
> Jos tekoälyjärjestelmän käytöstä voi aiheutua merkittävää haittaa moraalisille subjekteille eli moraalisten oikeuksien haltijoille, sen valmistajilla ja käyttäjillä on velvollisuus huolehtia siitä, että se on oikeintekijä.

Tämä periaate perustuu yleisempään käsitykseen välineellisen vahingoittamisen moraalisista implikaatioista. Jos hyväksymme edeltävät keinotekoisen toimijuuden vastaiset argumentit,

itseohjautuvat ja siten tietoteknisessä mielessä autonomiset järjestelmät ovat edelleen olennaisesti välineitä tai työkaluja. Jos aiheutan toiselle vahinkoa varomattomalla sahaamiselle, karanneella Roomballa, tai seonneella älyautolla, teen hänelle vääryyttä. Jos en ole kouluttanut koiraani kunnolla enkä pidä sitä lieassa sillä seurauksella, että se puree sinua, olen moraalisesti moitittava. Minun tehtäväni on huolehtia siitä, että koirani on vähintään ersatz-oikeintekijä, ja sama koskee robottiani. Toki tässä on komplikaatioita, koska kausaalinen vastuu robotin toiminnasta jakautuu eri tahoille, mutta jätän tämän kysymyksen tässä syrjään (kts. Hakli ja Mäkelä 2019).

On merkillepantavaa, että jos toteutan tämän velvollisuuteni, robottini läpäisee Allenin ja muiden (2000) esittämän moraalisen Turingin testin vertailevan version, jossa robotin arvioita erilaisista moraalisesti relevanteista skenaarioista verrataan ihmisten arvostelmiin. Tekoäly läpäisee testin, jos se on vähintään yhtä hyvä tässä tehtävässä kuin ihmiset keskimäärin. Toisin kuin Allen ja kumppanit esittävät, moraalinen Turingin testi ei siis sinänsä kerro moraalitoimijuudesta, vaan pelkästään oikeintekijyydestä.

Miten sitten voisimme rakentaa keinotekoisia oikeintekijöitä? Yksi houkutus on hyödyntää syväoppivien algoritmien kykyä löytää datasta säännönmukaisuuksia, mukaan lukien sellaisia, joita ihmiset eivät syystä tai toisesta sieltä löydä. Viimeaikaiset kilpaileviin algoritmeihin (GAN) perustuvat tekniikat ovat osoittaneet, että tekoäly kykenee luomaan myös uskottavia variaatioita, kuten keinotekoisia julkkiksia (Karras ja muut 2018). Sopivalla datalla ruokittu kone voisi siis oppia itse, mikä on oikein tai väärin. Mutta mitä dataa voisimme käyttää opettaaksemme koneen oikeintekijäksi? Emme informaatiota siitä, mitä ihmiset tosiasiassa tekevät, koska toimimme usein moraalittomasti, eivätkä moraaliperiaatteet ole kuvailevia. Emme myöskään dataa siitä, mitä pidämme moraalisena, koska olemme erehtyväisiä ja erimielisiä.

Toinen ilmeinen vaihtoehto on sisäänrakentaa järjestelmään joitakin moraalisia

periaatteita. Tätä edustavat esimerkiksi Isaac Asimovin (1950) kuuluisat robotiikan lait ja robootikko Ronald Arkinin (2009) ilmeisesti jossain määrin käytännössä toteuttama ”moraalinen ydin” autonomisille asejärjestelmille. Tämä ei välttämättä tarkoita paluuta vanhanaikaiseen tekoälyyn – ajatus on pikemminkin, että syväoppiviin järjestelmiin rakennetaan syötteen ja toiminnan väliin Arkinin kielellä ”pullonkaula”, jota epäeettiset toimintasuunnitelmat eivät läpäise. Jätän tekniseen toteutukseen liittyvät kysymykset toisille. Mutta mitkä periaatteet keinotekoiselle toimijalle tulisi ohjelmoida? Jotkut ovat ehdottaneet, että esimerkiksi itseohjautuvien autojen toimintaa vaaratilanteissa säätelevät säännöt voisi jättää yksittäisten käyttäjien valittavaksi. Mutta tämä on huono ehdotus. Vastausta siihen, pitääkö auton vaaratilanteessa suojella käyttäjää vai sivullisia, ei voi jättää käyttäjälle – hänen preferensseillään ei ole moraalista auktoriteettia oikean ja väärän toiminnan suhteen. Myöskään yleisesti hyväksyttyjä periaatteita ei pidä koneeseen ohjelmoida samasta syystä kuin aiemmin jo mainitsin – olemme erehtyväisiä ja erimielisiä. Filosofit eivät myöskään ole keskenään samaa mieltä, joten punnittua vastausta ei voi myöskään löytää kaupan hyllyltä. Nähdäkseni ainoa mielekäs lähtökohta on oikeiden periaatteiden ohjelmoiminen parhaan käsityksensä mukaan, mihin kuuluu aina vastuu ja riski väärässä olemisesta. Jos ohjelmoitkin keinotekoisen väärintekijän vaikka teit parhaasi, olet toiminut väärin, vaikka parhaasi tekeminen saattaakin olla anteeksiantoperuste.

Mutta entä jos ohjelmoija on kaikesta huolimatta epävarma oikeasta ratkaisusta? Silloin hänen on tehtävä valinta jonkinlaisen moraalisen epävarmuuden vallitessa. Keskustelu siitä, mitä tiedollinen epävarmuus moraalista merkitsee sille, mitä meidän tulee tehdä, on etiikan piirissä suhteellisen nuorta ja vakiintumatonta (kts. Bykvist 2017). Yksi konkreettinen esimerkki tästä on, että joku voi olla esimerkiksi 90% varma siitä, että lihansyönti on moraalisesti hyväksyttävää, mutta ajatella kuitenkin, että on 10% todennäköisyys, että se on vakavasti väärin. (Tarkat numeroarvot moraaliselle varmuudelle

ovat toki hieman kummallisia.) Jotkut filosofit ajattelevat, että näin uskovan tulee toimia sen käsityksen mukaan, josta hän on varmin. Mutta toiset pitävät tätä samanlaisena tilanteena kuin sitä, että on 90% varma että jos lähtee lumipyryssä ajamaan, ei törmää vastaantulevaan rekkaan, ja pitää 10% todennäköisenä sitä, että törmää. Näillä uskomusasteilla olisi ilmeisen vastuutonta lähteä tien päälle, jos vaihtoehto ei ole aivan kauhea. Hieman täsmällisemmin sanottuna, jos perille pääsemisen nettohyöty olisi vaikka 2 hyvinvointipistettä (käyttääkseni jälleen epärealistisen täsmällistä asteikkoa) ja törmäämisen nettohaitta -100, on helppo laskea, että mainituilla todennäköisyyksillä ajamisen odotushaitta ylittäisi sen odotushyödyn. Vastaavasti jos lihansyönnillä on minun kulinaariseen mielihyvääni perustuva matala positiivinen arvo moraalin näkökulmasta silloinkin jos se on hyväksyttävää, ja suuri negatiivinen arvo jos se on väärin (jolloin se olisi moraalisesti rinnastettavissa ihmisorjien syömiseen), maksimoimme moraalista odotusarvoa ryhtymällä kasvissyöjiksi, vaikka pitäisimme lihansyönnin vääryyttä varsin epätodennäköisenä.

Keskustelu toiminnasta moraalisen epävarmuuden vallitessa on tosiaankin varsin uutta, joten on vaikea sanoa, mitä siitä pitäisi ajatella. Kuvaamani valinta kasvissyönnin ja lihansyönnin välillä on siinä mielessä helppo, että toinen vaihtoehdoista on harvan mielestä väärin, joten on mahdollista pelata varman päälle. Mutta esimerkiksi itseohjautuvien autojen kohdalla vastaan tulee tilanteita, joissa kumpi tahansa vaihtoehto saattaa olla vakavasti väärin riippuen siitä, mitkä moraaliset periaatteet ovat oikeita. Kuvitellaan esimerkiksi, että neljää matkustajaa kuljettavan ajoneuvon sensorit havaitsevat yhtäkkiä kivivyöryn, jonka väistämiseksi on pakko ajaa kevyen liikenteen väylälle, jossa on jo pyöräilijä. Joidenkin käsitysten mukaan moraali vaatii tällaisessa tilanteessa yhden uhraamista monen pelastamiseksi (olettaen muun muassa, että kaikki ovat yhtä terveitä ja että heillä on suurin piirtein yhtä paljon elinvuosia jäljellä jne.), koska sen näkökulmasta meidän on edistettävä hyvinvointia puolueettomasti. Toisten mukaan taas tällaisessa tilanteessa olisi väärän loukata

ketään uhkaamattoman pyöräilijän perusoikeuksia siksi, että pelastaisi joukon ihmisiä, jotka ovat auton kyytiin noustessaan ottaneet riskin mahdollisen vaaratilanteen luomisesta toisille tai itselleen. Yksi tekniikan tarjoama mahdollisuus olisi ohjelmoida auto toimimaan tällaisessa tilanteessa jommallakummalla tavalla todennäköisyydellä, joka vastaa ohjelmoijan näihin vaihtoehtoisiin periaatteisiin kohdistuvan varmuuden astetta. Jos on vaikka 70% vakuuttunut hyvinvoinnin edistämisen olennaisuudesta ja 30% vakuuttunut oikeuksien kunnioittamisen tärkeydestä, voi siis ohjelmoida auton ajamaan vaaratilanteessa 70% todennäköisyydellä pyöräilijän päälle. Tämä maksimoisi (luontevin lisäoletuksin) moraalista odotusarvoa, kuten meidän pitäisi joidenkin mukaan tehdä silloin kun olemme epävarmoja. Mutta toisesta näkökulmasta tämä on absurdia. Jos ylipäänsä on moraalitotuuksia, jompikumpi periaatteista on epätosi – kuinka voisin toimia niin kuin pitää, jos otan ainakin merkittävän riskin siitä, että teen vakavaa vääryyttä jollekulle? Moraali ei vaadi nopanheittoa, vaan perusteltua valintaa.

Jätän kysymyksen epävarmuudesta tähän. Oma alustava kantani on, että meidän ei tulisi antaa robottien arpoa periaatteita, vaan ohjelmoida niille ainakin tietyt kiinteät perussäännöt, kuten ehdottomat kiellot joillekin teoille ja pisteytys joillekin itsessään hyville tai huonoille seurauksille (esimerkiksi tuskan aiheuttaminen tunteville olennoille on aina iso miinus). Mutta tämän ei tarvitse tarkoittaa täydellisen moraalikoodin ohjelmointia. Ainakin joissakin tapauksissa kiinteiden periaatteiden ohjelmointiin voisi yhdistää koneoppimisen niin sanotusta ”massojen viisaudesta”. Aristoteles sanoi aikanaan, että etiikassa ja filosofiassa ylipäänsä parhaan lähtökohdan harkinnalle muodostavat uskottavat näkemykset eli *endoksa*. Hänen mukaansa ”endoksaan kuuluvat mielipiteet, jotka hyväksyvät kaikki, suurin osa, tai viisaat – ja viisaiden joukossa kaikki tai suurin osa heistä, tai he, jotka ovat kaikkein huomattavimpia ja maineikkaimpia” (Aristoteles, *Topiikka* 100b21-33). Siinä määrin kuin moraalisia uskomuksia voi mallintaa koneen ymmärtämässä muodossa, niistä voisi kenties

suodattaa endoksaa, ehkä käyttämällä jotakin Googlen kuuluisan Pagerankin tapaista algoritmia, joka karkeasti sanoen antaa enemmän painoa niiden henkilöiden mielipiteille, joita muut paljon kuunnellut kuuntelevat. Tämä ei ratkaisisi erehtyväisyyden ongelmaa (jolle ei ylipäänsä ole olemassa ratkaisua), mutta se auttaisi erimielisyyden suhteen. Koneen voisi tähän tapaan ohjelmoida täydentämään ohjelmoijan itsensä asettamia sääntöjä oppimalla laajemmalta ihmisjoukolta.

Joissakin sovelluksissa olisi kenties mahdollista hyödyntää tekoälyjärjestelmien oppimiskykyä myös huomioimalla niiden omasta toiminnasta saatu palaute. Me ihmiset tajuamme joskus mokanneemme siitä, että toiset suuttuvat meille tai kohtelevat meitä kylmäkiskoisesti. Pelkästään se, että tekomme aiheuttavat jollekin pahan mielen voi olla syy kysyä itseltään, tuliko tehtyä jotain väärin. Koska kielteisten (ja toki myös myönteisten) tunteiden koneellinen tunnistaminen muun muassa ilmeistä, ruumiillisista reaktioista ja kirjoitustavasta kehittyy jatkuvasti (esimerkiksi Ghandeharioun ja muut 2017 diagnosoivat masennuksen melko tarkkaan ihon sähköaktiviteetin, unirytmien ja paikkatiedon avulla), esimerkiksi hoivarobotin voisi periaatteessa ohjelmoida muuttamaan toimintatapojaan tällaisen palautteen perusteella.

Mielenkiintoista kyllä, sentimentalistien kuten David Humen ja Adam Smithin (1759/2002) mukaan itse asiassa muodostamme moraalisäännöt osin juuri tällaisen palautteen perusteella. Meidän ”opetusdatamme” muodostuu karkeasti siitä, miten me itse reagoisimme johonkin tekoon toisten asemassa, tai miten reagoisimme jos olisimme puolueettomia mutta sympaattisia tarkkailijoita. Sympatian tai empatian kautta saamme siis palautetta tekojemme hyväksymisestä. Mutta sentimentalistit korostavat, että tällainen palaute täytyy ensin suodattaa – totta kai ihmiset joskus pahastuvat aivan ilman syytä, esimerkiksi erehtymisen, itsekkyyden tai kohtuuttomien odotusten takia. Siksi pelkästään tunnepalautteesta oppiminen ei riitä hyvään moraaliseen toimijuuteen, vaikka se voikin

toimia yhtenä syötteenä keinotekoisen oikeintekijän hienosäätämisessä.

4. Lopuksi

Jos meistä tekisi moraalisia pelkästään äly siinä mielessä kuin se on kyky ratkaista ongelmia oppimalla aiemmista onnistumisista ja epäonnistumisista, olisimme lähellä tilannetta, jossa keinotekoiset toimijat kykenisivät ihmisiä parempaan moraaliseen toimijuuteen. Mutta kuten olen esittänyt, itsenäinen moraalitoimijuus vaatii myös moraalista ymmärrystä, jonka edellytyksiä emme ymmärrä läheskään niin hyvin, että kykenisimme ohjelmoimaan niitä. Onkin viisaampaa pitää vastuu omissa käsissämme, ja pyrkiä rakentamaan pelkkiä oikeintekijöitä, eli koneita jotka toimivat huolellisesti punnitun käsityksemme mukaan mahdollisimman pitkälti siten, kuin ihmisten pitäisi toimia vastaavassa tilanteessa.

Lähteet

Allen, Colin, Varner, Gary & Zinser, Jason (2000) Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12:3, 251-261.

Aristoteles, *Topiikka*/Sofistiset kumoamiset. Gaudeamus, Helsinki.

Arkin, Ronald C. (2010). The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics* 9 (4):332-341.

Asimov, Isaac (1950). *I, Robot*. Gnome Press.

Audi, Robert (2004). *The Good in the Right: A Theory of Intuition and Intrinsic Value*. Princeton Up.

Baars, Bernard J. (1997). In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies* 4 (4):292-309.

Bykvist, Krister (2017). Moral uncertainty. *Philosophy Compass* 12 (3).

Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Coeckelbergh, Mark (2014). The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics. *Philosophy and Technology* 27 (1):61-77.

Cusk, Rachel. *Transit*. Jonathan Cape.

Darwall, Stephen L. (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard University Press.

Dennett, Daniel C. (1987). *The Intentional Stance*. MIT Press.

Doris, John M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Enoch, David (2014). A Defense of Moral Deference. *Journal of Philosophy* 111 (5):229-258.

Fischer, John Martin & Ravizza, Mark (1999). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Floridi, Luciano & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines* 14 (3):349-379.

Franklin, Stan (2003). Ida: A conscious artifact? *Journal of Consciousness Studies* 10 (4):47-66.

Grice, Paul 1974-75. Method in Philosophical Psychology (from the Banal to the Bizarre). *Proceedings and Addresses of the American Philosophical Association* 48: 23–53.

Grimm, Stephen R. (ed.) (2017). *Making Sense of the World: New Essays on the Philosophy of Understanding*. Oup Usa.

Gunkel, David J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology* 20 (2):87-99.

Hakli, Raul & Mäkelä, Pekka (2016). Robots, Autonomy, and Responsibility. In Johanna Seibt, Marco Nørskov & Søren Schack Andersen (eds.), *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*. Amsterdam, The Netherlands: IOS Press. pp. 145-154.

Hakli, Raul & Mäkelä, Pekka (2019). Moral Responsibility of Robots and Hybrid Agents.

Monist 102 (2): 259-275.

Hew, Patrick Chisan (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology* 16 (3):197-206.

Hills, Alison (2009). Moral testimony and moral epistemology. *Ethics* 120 (1):94-127.

Himma, Kenneth Einar (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11 (1):19-29.

Hume, David 1751. *An Enquiry Concerning the Principles of Morals*.

Jackson, Frank (1986). What Mary didn't know. *Journal of Philosophy* 83 (May):291-5.

Karras, Tero, Aila, Timo, Laine, Samuli ja Lehtinen, Jaakko (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *Proceedings of the International Conference on Learning Representations*.

Kauppinen, Antti (2015). Favoring. *Philosophical Studies* 172 (7):1953-1971.

Kauppinen, Antti (2017a). Empathy and Moral Judgment. In Heidi Maibom (ed.), *The Routledge Handbook of the Philosophy of Empathy*. Routledge.

Kauppinen, Antti (2017b). Sentimentalism, Blameworthiness, and Wrongdoing. In Karsten Stueber & Remy Debes (eds.), *Ethical Sentimentalism*. Cambridge University Press.

McDowell, John (1979). Virtue and Reason. *The Monist* 62 (3):331-350.

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

Parfit, Derek (2011). *On What Matters: Two-Volume Set*. Oxford University Press.

Picard, Rosalind W. (1995). Affective computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321.

Pylkkänen, Paavo (2007). *Mind, Matter and the Implicate Order*. Springer.

Ross, W. D. (1930). *The Right and the Good*. Clarendon Press.

Schlosser, Markus E. (2012). Taking Something as a Reason for Action. *Philosophical Papers* 41 (2):267-304.

Sharkey, Amanda (2017). Can robots be responsible moral agents? And why should we care?, *Connection Science*, 29:3, 210-216.

Smith, Adam (1759/2002). *A Theory of Moral Sentiments*. Toim. K. Haakonssen. Cambridge University Press.

Steiner, Adam P. & Redish, A. David (2014). Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nature Neuroscience*, 17, 995-1002.

Strawson, Peter F. (1962). Freedom and resentment. *Proceedings of the British Academy, Volume 48:* 1-25.

Torrance, Steve (2008). Ethics and consciousness in artificial agents. *AI and Society* 22 (4):495-521.

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

Wallach, Wendell & Allen, Colin (2008). *Moral Machines: Teaching Robots Right From Wrong*. Oxford University Press.

Westermarck, Edward (1906). *The Origin and Development of the Moral Ideas*. Freeport, N.Y., Books for Libraries Press.