Osaammeko rakentaa moraalisia toimijoita?

Kauppinen, Antti

Gaudeamus
2021

Kauppinen , A 2021 , Osaammeko rakentaa moraalisia toimijoita? julkaisussa P Raatikainen (Toimittaja) , Tekoäly, ihminen ja yhteiskunta : Filosofisia näkökulmia . Gaudeamus , Helsinki , Sivut 131-156 . < https://kauppa.gaudeamus.fi/sivu/tuote/tekoaly-ihminen-ja-yhteiskunta/3868931 >

http://hdl.handle.net/10138/340427

unspecified
publishedVersion

Osaammeko rakentaa moraalisia toimijoita?

Antti Kauppinen

Sinä ja minä olemme moraalitoimijoita. Karkeasti sanottuna kykenemme ottamaan huomioon moraalisia ja moraalisesti relevantteja seikkoja päättäessämme, mitä tehdä. Jos meillä ei olisi tällaista kykyä, toimisimme vielä nykyistäkin useammin väärin, eikä meitä voisi edes pitää siitä täysin vastuullisina. Moraalitoimijuus on siis tärkeää. Mitä se tarkalleen ottaen vaatii? Paul Grice esitti 1970-luvun puolivälissä yhden vaikutusvaltaisen menetelmän toimijuuden olennaisten piirteiden tunnistamiseksi. Hänen ajatuksensa 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.[1]

Gricen mielestä 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. 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 sekä 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 ottaisivat huomioon 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.[2] 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 oli ikinä tehnyt. Hän sanoi, että olemme keränneet valtavasti kielenkäyttöä ja tietoa elämästä, ja on voinut käydä niin, että valeinhimillisyydestä on tullut alkuperäistä syvempää ja suhteisiin valmiimpaa, että koneelta voi saada enemmän hellyyttä kuin kanssaihmiseltä. Mekanisoitu käyttöliittymähän tiivistää monta ihmistä, ei vain yhden.[3]

Cuskin kertojahahmon keskustelukumppani on toki tragikoominen hahmo, jota tosikkofilosofin on helppo syyttää projisoinnista ja antropomorfismista. Mutta vaikka emme olisi tosikkoja, monet näkevät, että keinotekoiseen moraalitoimijuuteen liittyy 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 pitkälle samalla tavoin kuin täysivaltaiset moraalitoimijat parhaimmillaan.

Moraalisesta toimijuudesta

Aloitetaan pohtimalla tarkemmin moraalisen toimijuuden käsitettä ja moraalisen toiminnan edellytyksiä. Ensin on otettava huomioon, että käytämme sanaa ”moraalinen” ja sen sukulaisia sekä kuvailevasti että arvottavasti. Kuvailevassa mielessä moraalisia toimijoita ovat ne, jotka kykenevät muodostamaan jonkinlaisen käsityksen oikeasta ja väärästä ja toimimaan sen mukaisesti. Arvottavassa mielessä moraalisia ovat ne, jotka toimivat (ainakin enimmäkseen) oikein tai hyvin. Niinpä esimerkiksi juutalaisten joukkotuhoa organisoinut Adolf 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 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. Nämä representaatiot puolestaan aiheuttavat ruumiin- tai mielenliikkeitä. Esimerkiksi rotat kykenevät laboratoriokokeiden valossa jopa astetta vaativampaan toimijuuteen. Ne kykenevät toimimaan *suunnitelmallisesti* eli representoimaan ajallisesti etäisempiä päämääriä ja monivaiheisia keinoja

niiden saavuttamiseen sekä pidättäytymään muiden halujen tyydyttämisestä silloin, kun se haittaisi päämäärien saavuttamista. Suunnitelmallisuus edellyttää siis jonkinasteista *itsehallintaa* ja itsehillintää. Aivotutkimusten mukaan rottienkin on väitetty muun muassa tuntevan katumusta, jos ne sortuvat vähempiarvoisiin pikavoittoihin.[4]

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 niiden saavuttamiseksi) ja vielä tehtävä se tavalla, joka ottaa huomioon moraalin vaatimukset. (Kun puhumme kuvailevassa mielessä, kyse on vaatimuksista, jotka ovat toimijan käsityksen mukaan moraalisia – ne eivät siis välttämättä ole aidosti moraalisia.) Lisäksi tällä arvioinnilla on oltava oikeanlainen merkitys sille, mitä toimija tekee.

Kun ryhdymme määrittelemään moraalitoimijuutta tarkemmin, lienee siis hedelmällistä aloittaa seikoista, jotka moraalin näkökulmasta puoltavat tietynlaista toimintaa eli siis moraalisista perusteista. Esimerkiksi se, että lapsen tönäiseminen kiireessä pois tieltäni tuottaisi vaaratilanteen viattomalle, on peruste olla tönäisemättä häntä. Voimme myös sanoa, että lapselle aiheutuva vaara *antaa* perusteen olla tönäisemättä.[5] Tämä puhetapa on siitä hyödyllinen, että se auttaa erottamaan kaksi tosiasiaa, yhtäältä sen empiirisen faktan, että tietty teko aiheuttaa lapselle vaaran, 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.[6] On väitetty esimerkiksi, että ihmisen nimi vaikuttaa tilastojen valossa hänen ammatinvalintaansa tai että jos ihmistä huomaamatta viritetään (engl. *priming*) ajattelemaan tiettyjä asioita, hänet saa esimerkiksi pidättäytymään epärehellisyydestä. On tällä hetkellä empiirisesti kiistanalaista, kuinka merkittäviä nämä tiedostamattomat vaikutukset ovat, koska useita keskeisiä tuloksia ei ole kyetty toisintamaan. Oli miten oli, toinen tapa, jolla moraaliset seikat voivat vaikuttaa tekoihimme, on, että kohtelemme jotain asiaa riittävänä perusteena toiminnalle ja toimimme sen mukaisesti. Joskus, tosin suhteellisen harvoin, punnitsemme perusteita harkitusti. Kenties mietin aamulla, minkä takin laittaisin päälleni, ja katsoessani lämpömittaria totean, että tällä pakkasella on hyvä syy laittaa ulsteri. Olisi kuitenkin virhe luulla, että jonkin asian perusteena pitäminen vaatii eksplisiittistä harkintaa. Kadulla kiirehtiessäni en missään vaiheessa tietoisesti pysähdy miettimään, tönäisenkö lasta, mutta voin silti pitää lapselle koituvaa vaaraa perusteena olla tönimättä heitä. Tämä näkyy taipumuksissa, jotka toteutuisivat todenvastaisissa tilanteissa.[7] Jos esimerkiksi vahingossa tönäisisin jotakuta, pyytäisin anteeksi; jos sinä tönäisisit minua, saattaisin tölväistä sinua. Lapselle aiheutuva haitta ei ole vain (kausaalinen) syy sille, miksi en tönäise, vaan myös minun itseni hyvänä pitämä syy sille.

Aiheemme kannalta on olennaista erottaa nimenomaan *moraalisena* perusteena kohteleminen pelkästään perusteena kohtelemisesta. Nähdäkseni moraaliset perusteet liittyvät ennen kaikkea niin sanottuihin *reaktiivisiin asenteisiin*, kuten suuttumukseen, halveksuntaan, kiitollisuuteen, syyllisyyteen, häpeään ja ylpeyteen. Kuten esimerkiksi Edward Westermarck, Peter Strawson ja Stephen Darwall ovat korostaneet, moraalisen ajattelun keskeinen funktio on nimenomaan tällaisten asenteiden ohjaaminen ja ilmaiseminen ja sitä kautta toimintaan vaikuttaminen.[8] Jos esimerkiksi näen jonkun toimivan tavalla, jota vastaan on mielestäni riittävät moraaliset perusteet, olen taipuvainen

asennoitumaan häneen kielteisesti tai ainakin pitämään kielteisiä reaktiivisia asenteita paikallaan olevina. 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 ottamaan ne oikein huomioon toiminnassamme. Tämä vaatii kuitenkin edellisten erottelujen valossa selvennystä. Erottelujen 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. Hänellä on pakkomielle päästä demoneista eroon.
– Ben tönäisee lapsen kadulle, koska hän luulee virheellisesti, että se on lapsesta kivaa. Ben ei ajattele, että tönäiseminen voisi olla vaarallista.
– Cersei tönäisee lapsen kadulle, koska hänestä meillä on riittävän hyvä syy satuttaa lapsia.
– Desi jättää tönäisemättä lapsen kadulle, koska hänellä on tiedostamaton taipumus vältellä toisten satuttamista.
– Elina jättää tönäisemättä lapsen kadulle, koska lapsen takki on luminen, eikä hän halua kastella kalliita nahkahanskojaan.
– 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 toimii oikein väärästä syystä. Vain Frank tekee oikein oikeasta syystä.

Kutsun kolmea ensimmäistä toimijaa *väärintekijöiksi* ja kolmea jälkimmäistä *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 voida pitää moraalisesti vastuullisena toimijana. Ben ja Cersei sen sijaan todennäköisesti kykenevät toimimaan oikein, jos he vain käyttävät asianmukaisesti kykyjään muodostaa empiirisiä tai moraalisia uskomuksia. He ovat siten moraalisesti vastuussa virheestään, ellei niille ole jotain anteeksiantoperustetta (kuten omasta laiminlyönnistä johtumatonta tietämättömyyttä tai ulkoista pakotusta).

Aiheemme näkökulmasta meitä kiinnostavat erityisesti erityyppiset *oikeintekijät*. (Jostain syystä tällainen sana puuttuu ainakin kaikista tuntemistani kielistä.) Voi olla, että tavalliset moraalitoimijat toimivat toisinaan moraalisesti hyväksyttävällä tavalla syistä, joita he eivät tiedosta. Entäpä jos Desi tekee säännöllisesti 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. Hän ei esimerkiksi koskaan vetoa siihen oikeuttaakseen toimintaansa eikä reagoi kielteisesti niihin, jotka jättävät moraalisesti relevantin seikan huomiotta. Jos hän ei kykene kohtelemaan perustetta perusteena, 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 luultavasti moraalitoimija, joka tekee oikein väärästä syystä. Näin sanoessamme oletamme, että hän kykenisi tekemään toisinkin ja vas-

taamaan 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 Immanuel Kant korosti, on merkittävä ero sen välillä, toimimmeko moraalin vaatimusten *mukaisesti* vai niiden *ohjaamina*. Moraalitoimijat ainakin 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. Mikä sitten on leimallista moraalisille perusteille ja säännöille? Kysymyksemme ei koske *aitoja* perusteita tai *oikeita* sääntöjä, joten olisi virhe painottaa vastauksessa tiettyjä sisältöjä, kuten toisten edun huomioon ottamista. 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 moraalitoimija jättää teon tekemättä, vaikka siitä olisi hänelle hyötyä, tai sitten tuntee syyllisyyttä teosta, joka ei ollut moraalin mukainen. Lyhyesti: 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 David 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 jakavan. Hänen täytyy siten nousta yksityisen ja erityisen tilanteensa yläpuolelle ja valita näkökulma, joka on yhteinen hänelle ja toisille.[9]

Hume luonnehtii tässä asiaa sentimentalistisessa sävellajissa, mutta yleisemmin muotoiltuna ajatuksen hyväksyvät usean eri metaeettisen koulukunnan edustajat.

Moraalitoimijuudeksi ei tietenkään riitä, että kykenemme moraalisiin arvostelmiin. Ne on myös kyettävä panemaan toimeen. Se edellyttää suunnitelmallisuutta ja itsehallintaa. Nämä seikat korostuvat siksi, että moraaliseksi koetut säännöt usein vaativat luopumaan jostain, mitä muuten haluaisimme.

Moraalitoimijuudesta on luontevaa erottaa eri asteita. On *minimaalisia* moraalitoimijoita, jotka omaksuvat periaatteensa ja asenteensa toisilta mutta eivät kykene kyseenalaistamaan oppimaansa. He ajattelevat kyllä moraalisesti ja toimivat enemmän tai vähemmän käsitystensä mukaisesti, mutta he joutuvat ymmälleen, jos normeilla on ristiriitaisia implikaatioita, jos niitä pitää soveltaa luovasti tai jos joku haastaa heidät puolustamaan niitä. Esimerkiksi lapset ja joidenkin arvioiden mukaan jotkut autismin kirjolla olevat ihmiset kuuluvat tähän ryhmään.[10] Minimaalisten moraalitoimijoiden 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. Oikein tekemisellä on kuitenkin heille väliä.

Sen sijaan *täysivaltaiset* moraalitoimijat kykenevät *ymmärtämään*, miksi jotkin teot ovat väärin. Tämä ymmärrys voi myös motivoida heitä asianmukaisesti. Täysivaltaiset moraalitoimijat kykenevät muodostamaan itsenäisiä arvostelmia eivätkä siten ole riippuvaisia toisista moraalisten ratkaisujen suhteen. Tekojen moraalisuuden ymmärtämiseen sisältyy muun muassa

kokonaisuuden eri osien roolin ja niiden välisten riippuvuussuhteiden hahmottaminen. Tämä mahdollistaa vastaamisen kysymyksiin siitä, minkä pitäisi muuttua teon moraalisen statuksen muuttamiseksi.[11] Nyrkkisääntönä on esimerkiksi väärin kertoa rasistisia vitsejä. Tällä lienee tekemistä sen kanssa, että se ilmentää ansaitsematonta ylenkatsetta, ulkopuolistaa pilkan kohteita ja usein vahvistaa heihin kohdistuvia ennakkoluuloja. Rasistisissa vitseissä on kyse eriarvoisesta kohtelusta, joka antaa aihetta loukkaantua. Mitä vikaa on eriarvoisessa kohtelussa tai ylenkatseessa tai loukkaavassa käytöksessä ja miksi? Miksi nämä seikat puhuvat tekoa vastaan? Onko joskus kenties 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 karkeasti erottaa kolme filosofista koulukuntaa sen suhteen, mitä moraalinen ymmärrys vaatii. *Intuitionistit* ajattelevat, että kykenemme ymmärryksellä välittömästi tavoittamaan moraalisia totuuksia, joko yleisiä tai erityisiä. Esimerkiksi W. D. Rossin ja Robert Audin mukaan on itsestään selvää, että lupaukset on lähtökohtaisesti pidettävä ja vahingon tekemistä vältettävä.[12] Kuka tahansa, joka ymmärtää nämä moraaliset väitteet riittävän hyvin, 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, 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. Viimeisimpänä, muttei suinkaan vähäisimpänä, *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 jonkin teon ymmärtäminen vääräksi motivoi olemaan tekemättä sitä, sillä 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 kuuluisaa ajatuskoetta[13] mukaillen esimerkiksi Mary, joka ei ole koskaan kokenut nälkää eikä mitään muutakaan epämiellyttävää asiaa. Mary voi kyllä kuulopuheiden perusteella ymmärtää, että nälkä hankaloittaa muita toimintoja, kuten vaikkapa keskittymistä. Hänen on myös mahdollista *tietää*, että se on itsessään huono asia, jos jokin luotettava taho kertoo niin.[14] Mutta kun Mary ensimmäistä kertaa itse näkee nälkää, hän vaikuttaisi oppivan jotain siitä, miltä nälkä tuntuu, mutta 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 syyn sille vasta, kun joutuu itse uhrin asemaan.

Vaikka olenkin yleisesti ottaen sentimentalisti,[15] en tässä ota kantaa siihen, mikä näistä malleista on paras. Jotta moraalitoimijuuden merkitys selkiytyisi, vertaan vielä kahta erilaista oikeintekijää, joista toinen ei ole lainkaan moraalitoimija ja toinen on täysivaltainen sellainen. Ensimmäinen voi olla vaikkapa psykopaatti, joka ei kirjaimellisesti kykene ymmärtämään moraalin merkitystä 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. Luettuaan Thomas Hobbesia hän voi myös haluta kaikkien muidenkin noudattavan näitä sääntöjä, koska siitä on ei-moraalisia hyötyjä. Tästä syystä hän on valmis rankaisemaan väärintekijöitä. Jos vielä oletamme, että hänelle satutaan opettamaan juuri oikeat säännöt, hän saattaa olla vir-

heetön 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 monesti oikein, vaikka hänet olisi opetettu tekemään väärin, vaikka toiset tosiasiassa tekisivät väärin ja vaikka hänen yleinen motivaatiotilansa olisi erilainen. Näin on siksi, että hänen motivaationsa oikein tekemiseen juontaa ymmärryksestä itsestään. On siis monta syytä, miksi on tärkeää erotella, olemmeko täysivaltaisia moraalitoimijoita, vaikka kaikissa tilanteissa ero ei näy tosiasiallisessa toiminnassa.

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 nyt syrjään yleisemmät huolenaiheet siitä, kykenevätkö koneet suuntautumaan maailmaan ja aidosti tavoittelemaan päämääriä (näitä seikkoja käsittelee esim. Searlen kritiikki[16]). Vaikka filosofien esittämät huolenaiheet ovat varteenotettavia, on joka tapauksessa mielenkiintoista pohtia, liittyykö keinotekoiseen moraaliseen toimijuuteen *erityisiä* ongelmia. Aloitan täyttä moraalitoimijuutta vastaan esitetyistä argumenteista. Ensimmäinen niistä vetoaa vapaaseen tahtoon:[17]

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.

En käytä tämän argumentin tarkasteluun paljoa aikaa. Moraalitoimijuuden yhteys vapaaseen tahtoon kulkee vastuun kautta, joten jätän tässä syrjään syvälliset kysymykset vapaan tahdon mahdollisuudesta ja todellisuudesta[18] ja keskityn vastuun itsensä edellytyksiin. Yksi perustelu tälle on se, ettei ole lainkaan harvinaista kiistää vastuun edellyttävän vapaata tahtoa (ja siten hylätä argumentin premissi 2). Näin tekevät muun muassa John Martin Fischer ja Mark Ravizza, jotka ovat valmiita hyväksymään, ettei meillä ole vapaata tahtoa ainakaan kovin vahvassa mielessä. Heidän mielestään voimme siitä huolimatta olla moraalisesti vastuullisia karkeasti silloin, kun tekomme ovat seurausta sellaisista järkiperusteille herkistä mielen kyvyistä, joiden ansiosta olisimme *voineet* toimia oikein niissäkin tilanteissa, joissa tosiasiassa teemme väärin.[19]

Toinen argumentti[20] 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 sekä Raul Hakli ja Pekka Mäkelä ovat hieman eri tavoin esittäneet, että esiohjelmointi vastaa moraalisesti tällaista manipulaatiota.[21] 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, jonka työhön he nojaavat, korostaa, että manipuloidut tai esiohjelmoidut arvot ovat ”käytännöllisesti lukkoon lyötyjä” – toimijat eivät kykene rationaalisesti ja vapaaehtoisesti muuttamaan niitä, koska tämä edellyttäisi jonkinlaista pääsyä niiden ulkopuolelle.[22]

Tämä on lupaava argumentti, mutta en kuitenkaan pidä sitä yksin ratkaisevana. On ensinnäkin oltava tarkkoja siitä, ettei moraalitoimijuuden rima nouse niin korkealle, etteivät useimmat aikuiset ihmisetkään ylitä sitä. Toiseksi, ainakin Hakli ja Mäkelä ovat valmiita olettamaan, että synkroniset (toimijalla tiettynä ajankohtana olevat) 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 erivärisiä 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 mahdollinen argumentti sitä vastaan:[23]

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, mutta jätin vielä avoimeksi, mikä on paras malli moraalisesta ymmärryksestä tai kompetenssista. Sentimentalismin mukaan olion on karkeasti sanottuna kyettävä empatiaan ja tunteiden hallintaan laajemmasta perspektiivistä, jotta kyseinen olio voi ymmärtää moraalin jujun.[24] Lienee varsin uskottavaa, että tällaisen ymmärryksen keinotekoisesta tuottamisesta ollaan kaukana. Yksi syy skeptisyyteen tämän suhteen on siinä, ettei keinotekoisten tunteiden tuottamisessa ole edistytty lainkaan, vaikka keinoäly on muuten ottanut valtavia edistysaskeleita. (Palaan tähän kohta.) Joku voisi ajatella, että samasta syystä rationalistinen malli antaa syytä toiveikkuuteen. 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[25] 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”[26] ja joskus hyveelliseen ”näkemiseen”, jolle ei voi antaa arvovapaita kuvailevia kriteerejä.[27] Tätä tietynlaista epämääräisyyttä voi toki pitää näiden näkemysten filosofi-

sena ongelmana, mutta ainakin se viittaa siihen, että moraalinen kompetenssi vaatii juuri sellaista laajakatseista arvostelukykyä, jonka toteuttaminen vähemmänkin vaativissa kysymyksissä on osoittautunut erityisen vaikeaksi, vaikka affektiivisuuden haasteet jätettäisiin syrjään.

Olen tähän mennessä puhunut täysivaltaisen moraalitoimijuuden toteuttamisen kompastuskivistä. Entä suppeammin ymmärretty toimijuus, jonka tärkeimpänä edellytyksenä on ylipäänsä joihinkin asioihin suhtautuminen leimallisesti moraalisella tavalla ja sen mukaisesti toimiminen? Keskityn tässä vain yhteen mahdolliseen 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 (kokemustietoisuutta eli fenomenaalista tietoisuutta).
4. Keinotekoisilta toimijoilta puuttuu kokemustietoisuus.
5. Siis, keinotekoiset toimijat eivät voi olla moraalitoimijoita.

Olen puolustanut edempänä kahta ensimmäistä premissiä, ja moni kenties hyväksyy neljännen. Entä kolmas? Kaikki eivät ole siitä yksimielisiä. Wendell Wallach ja Colin Allen 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.*[28]

Ydinajatuksena tässä on nähdäkseni se, että moraalitoimijuuden edellytykset täyttyvät, 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.[29] Tutkimus on paljolti keskittynyt ensimmäiseen näistä. Esimerkiksi Stan Franklinin luoma LIDA-malli (engl. *learning intelligent distribution agent*) lähtee Bernard Baarsin ajatuksesta, jonka mukaan tietoisuus on eräänlainen ”globaali työtila”, jonka hallinnasta tiedostamattomien prosessien yhteenliittymät kilpailevat.[30] 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 tärkeystasolle luokitellun informaation yleiseen työtilaan, josta se välittyy kaikille aliohjelmille. Franklin esittää, että tietoisuuden funktio on esimerkiksi auttaa toimimaan uusissa tilanteissa, varoittaa vaaroista, kertoa toimintamahdollisuuksista ja mahdollistaa ympäristön piirteisiin sopiva käyttäytyminen. Hän uskoo, että hänen mallinsa toteuttava järjestelmä toimii näin.[31]

Tämä herättää paljon kysymyksiä, jotka liittyvät niin sanottuun tietoisuuden vaikeaan ongelmaan eli sen selittämiseen, kuinka fysikaaliset prosessit voivat aiheuttaa tai toteuttaa asioiden subjektiivisen tunnun.[32] Tässä yksi pikainen argumentti kokemustietoisuuden irrelevanssia vastaan:

1. Jos on mahdollista jäljitellä jäänteettä tietoisuutta funktionaalisesti ilman fenomenaalista tietoisuutta, fenomenaalinen tietoisuus on *epifenomenaalista*. (Toisin sanottuna, zombilta 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. Sellaiseksi se vaikuttaa kuitenkin 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 funktio liittyy siihen, mitä joskus kutsutaan alkuperäiseksi intentionaalisuudeksi, eli sellaiseen maailmaan suuntautumiseen, joka ei edellytä minkäänlaista tulkintaa tai roolia kokonaisuuden toiminnassa.[33] Tästä seuraa tietysti, etteivät zombit voi käyttäytyä täsmälleen samoin kuin vastaavassa tilanteessa olevat kokemustietoiset yksilöt. Tieteellistä näyttöä tästä ei ole, mutta elokuva- ja televisiozombit vaikuttavat kyllä varsin kylmäkiskoisilta!

Kritiikin voi torjua ehdottamalla, että saman funktion voi yleisesti ottaen toteuttaa eri tavoilla. Kuten Wallach ja Allen huomauttavat, Deep Blue voitti Kasparovin pelaten aivan eri periaatteilla kuin shakkia taitava ihminen. Miksei myös tunteiden ja kokemustietoisuuden moraalista funktiota voisi korvata toisella tavalla toimivalla järjestelmällä?[34] Tämä olisi kenties mahdollista, jos niillä olisi vain välineellinen rooli. Mutta siinä määrin kuin tunteet ovat *olennaisesti konstitutiivisia* välittämiselle ja asioiden moraalisena kohtelemiselle, niitä ei voi korvata jollakin muulla arkkitehtuurilla.

Ongelman voisi kenties kiertää, jos keinotekoiselle toimijalle voisi luoda keinotekoiset tunteet. Affektiivinen tietojenkäsittelytiede pyrkii opettamaan tekoälysovelluksia tunnistamaan tunteita ja vastaamaan niihin soveliaasti. Alan tutkimus ei kuitenkaan tavoittele synteettisten tunteiden luomista sinänsä.[35] Monet

pitävät tehtävää mahdottomana. Esimerkiksi Steve Torrance ja Amanda Sharkey argumentoivat, että tunteet ja ylipäänsä kokemuksellisuus ovat olennaisesti ruumiillisten, itseorganisoituvien biologisten organismien ominaisuuksia.[36]

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 lähtevät siitä, että voimme tarkastella järjestelmiä eri abstraktiotasoilla. Abstraktiotason valinta vaikuttaa heidän mukaansa siihen, ovatko järjestelmät moraalisia toimijoita vai eivät.[37] Sopivalla abstraktiotasolla myös tekoälyjärjestelmät täyttävät heidän ehtonsa toimijuudelle, koska ne vuorovaikuttavat ympäristönsä kanssa sisäisten tilojensa ohjaamina ja voivat muuttaa reaktioitaan ohjaavia sääntöjä. Mikäli tekoälyjärjestelmien 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” (engl. *accountable*) esimerkiksi siinä mielessä, että jos ne toimivat väärin, ne kannattaa korjata tai purkaa. Mark Coeckelberghin ja David Gunkelin mielestä meidän tulisi lähestyä moraalitoimijuutta ja -subjektiutta relationaalisesti ja aloittaa siitä, mitä asioita kohtelemme moraalisina niiden sisäsyntyisistä määreistä riippumatta.[38] Coeckelbergh tiivistää, että ”tässä lähestymistavassa ei enää ole kuilua ’oikean’ tavan nähdä robotti ja minun ’havaintoni’ robotista välillä.”[39]

Nämä yritykset ovat monella tapaa ongelmallisia. Kummatkin suhteellistavat toimijuuden joko intresseihimme tai suhtautumistapoihimme. On toki totta, että voi olla joitakin tarkoituksia varten hyödyllistä puhua erilaisista järjestelmistä toimijuuden kielellä, kuten Daniel Dennett jo kauan sitten huomautti.[40] On kuitenkin ongelmallista, 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. Moraalitoimijuus edellyttää ainakin vastuun kannalta välttämättömiä kykyjä. Se, mistä Floridi ja Sanders puhuvat, ei tosiasiassa ole missään mielessä moraalinen tilivelvollisuus vaan ainoastaan kausaalinen vastuu. Jos järjestelmä toimii epätoivottavasti, on olennaista kyetä tunnistamaan, mitä siinä pitää muuttaa. Puhe tilivelvollisuudesta kuitenkin vain sumentaa asiaa. Cockelberghin ja Gunkelin kriteerit taas ovat niin väljiä, että niiden mukaan lasteni pehmolelut olisivat ilmeisesti moraalisia subjekteja. Jos Cockelbergh ja Gunkel 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.

Kohti keinotekoisia oikeintekijöitä

Olen esittänyt, että keinotekoisten moraalitoimijoiden luominen on monin tavoin haastavaa. Sanomani ei kuitenkaan ole pelkästään negatiivinen. Hyvä moraalitoimija on takuuvarmempi kuin pelkkä oikeintekijä, koska itsenäinen moraalinen ymmärrys tekee hänestä vähemmän herkän ohjelmoinnin tai kasvatuksen virheellisille lähtöoletuksille ja paremman arvioimaan moraalin vaatimuksia uusissa ja ennakoimattomissa tilanteissa. Olisi silti valtava kehitysaskel, jos pystyisimme luomaan tavallisissa oloissa luotettavia keinotekoisia oikeintekijöitä.[41] Tällöin koneet toimivat moraalisesti oikealla tavalla tai ainakin samoin kuin moraalitoimija vastaavassa tilanteessa. Tätä viittausta moraalitoimijan kaltaiseen toimintaan tarvitaan, koska ei ole kiistatonta, voimmeko ylipäänsä mielekkäästi sanoa, että jonkun *pitää* tehdä jotain, jos

hän ei voi olla tilivelvollinen mistään.[42] 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 moraalisesta merkityksestä. 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 sahaamisella, karanneella Roomballa tai seonneella älyautolla, teen hänelle vääryyttä. Jos en ole kouluttanut koiraani kunnolla enkä pidä sitä hihnassa 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.[43]

On merkillepantavaa, että jos toteutan tämän velvollisuuteni, robottini läpäisee Allenin ja muiden 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.[44] 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.[45] Sopivalla datalla ruokittu kone voisi siis periaatteessa oppia itse, millaiset asiat ovat oikein tai väärin. Mitä dataa voisimme käyttää opettaaksemme koneen oikeintekijäksi? Tähän ei sovi informaatio siitä, mitä ihmiset tosiasiassa tekevät, koska toimimme usein moraalittomasti, eivätkä moraaliperiaatteet ole kuvailevia. Emme myöskään voi käyttää dataa siitä, mitä *pidämme* moraalisena, koska olemme erehtyväisiä ja erimielisiä.

Toinen ilmeinen vaihtoehto on sisäänrakentaa järjestelmään joitakin moraalisia periaatteita. Tätä edustavat esimerkiksi Isaac Asimovin kuuluisat robotiikan lait[46] ja robootikko Ronald Arkinin ilmeisesti jossain määrin käytännössä toteuttama ”moraalinen ydin” autonomisille asejärjestelmille.[47] 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. Tekniset kysymykset ovat asia erikseen, mutta mitkä periaatteet keinotekoiselle toimijalle tulisi ohjelmoida? Jotkut ovat ehdottaneet, että esimerkiksi itseohjautuvien autojen toimintaa vaaratilanteissa säätelevät säännöt voisi jättää yksittäisten käyttäjien valittavaksi. Tämä on huono ehdotus. Vastausta siihen, pitääkö auton vaaratilanteessa suojella käyttäjää vai sivullisia, ei voi jättää käyttäjälle itselleen, koska hänellä ei ole moraalista auktoriteettia ratkaista kysymyksiä toisiin mahdollisesti kohdistuvan haitan oikeutuksesta. Koneeseen ei pidä myöskään ohjelmoida yleisesti hyväksyttyjä periaatteita samasta syystä kuin aiemmin jo mainitsin – olemme erehtyväisiä ja erimielisiä. Nähdäkseni paras lähtökohta on oikeiden periaatteiden ohjelmoiminen parhaan käsityksen mukaan – ohjelmoija ei voi ulkoistaa niiden valintaa kenellekään, kunhan toimii lain määrittelemissä rajoissa. Tässä on kuitenkin otettava huomioon eettisten ajattelijoiden esittämät perustellut mielipiteet (mistä lisää hetken kuluttua) ja jätettävä

syrjään ohjelmoijan omaan etuun liittyvät perusteet (kuten se, että käyttäjän etusijalle asettavat järjestelmät todennäköisesti myyvät paremmin). Ohjelmointiin kuuluu joka tapauksessa aina vastuu ja riski väärässä olemisesta. Jos vahingossa päätyykin ohjelmoimaan keinotekoisen väärintekijän, on tehnyt väärin – vaikka parhaansa tekeminen saattaakin olla anteeksiantoperuste.

Entä jos ohjelmoija on kaikesta huolimatta epävarma oikeasta ratkaisusta? Silloin hänen on tehtävä valinta jonkinlaisen moraalisen epävarmuuden vallitessa. Etiikan piirissä tähän liittyvä keskustelu on suhteellisen nuorta ja vakiintumatonta.[48] Yksi konkreettinen esimerkki valinnasta moraalisen epävarmuuden vallitessa on, että joku voi olla esimerkiksi 90-prosenttisen varma siitä, että lihansyönti on moraalisesti hyväksyttävää, mutta ajatella kuitenkin, että on 10 prosentin todennäköisyys sille, että se on vakavasti väärin. (Tarkat numeroarvot moraaliselle varmuudelle ovat toki ylipäänsä hieman kummallisia.) Jotkut filosofit ajattelevat, että näin uskovan tulee toimia sen käsityksen mukaan, josta hän on varmin. Toiset sen sijaan pitävät tätä samanlaisena tilanteena kuin sitä, että on 90-prosenttisen varma, että jos lähtee lumipyryssä ajamaan, ei törmää vastaantulevaan rekkaan, ja pitää 10-prosenttisesti 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 ajamaan lähtemisen odotushaitta ylittäisi sen odotushyödyn. Vastaavasti jos lihansyönnillä on minun kulinaariseen mielihyvääni perustuva matala positiivinen arvo moraalin näkökulmasta silloinkin, kun se on hyväksyttävää, ja suuri negatiivinen arvo, jos se on väärin (jolloin se olisi moraalisesti rinnastettavissa ihmisorjien syömiseen), maksimoin moraalista odotusarvoa ryhtymällä kasvissyöjiksi, vaikka pitäisin 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. Tietyissä tilanteissa on siis mahdollista pelata varman päälle. Esimerkiksi itseohjautuvat autot joutuvat kuitenkin tilanteisiin, joissa kumpi tahansa vaihtoehto saattaa olla vakavasti väärin. 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 väärin loukata ketään uhkaamattoman pyöräilijän perusoikeuksia siksi, että pelastaisi joukon ihmisiä, jotka ovat auton kyytiin noustessaan ottaneet riskin mahdollisen vaaratilanteen luomisesta toisille tai itselleen.[49] 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-prosenttisen vakuuttunut hyvinvoinnin edistämisen olennaisuudesta ja 30-prosenttisen vakuuttunut oikeuksien kunnioittamisen tärkeydestä, voi siis ohjelmoida auton ajamaan vaaratilanteessa 70 prosentin todennäköisyydellä pyöräilijän päälle. Tämä maksimoisi (luontevin lisäoletuksin) moraalista odotusarvoa, kuten meidän pitäisi joidenkin mukaan tehdä silloin, kun olemme epävarmoja. Toisesta näkökulmasta tämä on absurdia. Jos ylipäänsä on olemassa 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). Tämän ei kuitenkaan 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 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”.[50] 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 ottamalla huomioon niiden omasta toiminnasta saatu palaute. Joskus me ihmiset tajuamme mokanneemme, kun toiset suuttuvat meille tai kohtelevat meitä kylmäkiskoisesti. Pelkästään se, että tekomme aiheuttavat jollekin pahan mielen, voi olla syy kysyä itseltämme, tuliko tehtyä jotain väärin. Koska kielteisten (ja toki myös myönteisten) tunteiden koneellinen tunnistaminen muun muassa ilmeistä, ruumiillisista reaktioista ja jopa ihmisen tavasta kirjoittaa kehittyy jatkuvasti,[51] esimerkiksi hoivarobotin voisi periaatteessa

ohjelmoida muuttamaan toimintatapojaan tällaisen palautteen perusteella.

Mielenkiintoista kyllä, sentimentalistien, kuten David Humen ja Adam Smithin[52], mukaan me itse asiassa muodostamme omat moraaliset periaatteemme osin juuri tällaisen palautteen perusteella. Meidän ”opetusdatamme” muodostuu karkeasti siitä, miten me itse reagoisimme johonkin tekoon toisten asemassa, tai siitä, miten reagoisimme, jos olisimme puolueettomia mutta sympaattisia tarkkailijoita. Sympatian tai empatian kautta saamme siis palautetta tekojemme hyväksymisestä. Sentimentalistit kuitenkin 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.

Lopuksi

Jos pelkästään äly siinä mielessä kuin se on kyky ratkaista ongelmia oppimalla aiemmista onnistumisista ja epäonnistumisista tekisi meistä moraalisia, olisimme lähellä tilannetta, jossa keinotekoiset toimijat kykenisivät jopa ihmisiä parempaan moraaliseen toimijuuteen. Kuten olen esittänyt, itsenäinen moraalitoimijuus vaatii kuitenkin 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.

1. Grice 1974–1975, 37.
2. Arkin 2010.
3. Cusk 2016, oma käännökseni.
4. Steiner & Redish 2014.
5. Vrt. Parfit 2011, 32.
6. Doris 2015.
7. Schlosser 2012; Kauppinen 2015.
8. Westermarck 1906; Strawson 1962; Darwall 2006.
9. Hume 1983, oma käännökseni.
10. Ks. Kauppinen 2017a.
11. Hills 2009; Grimm 2017.
12. Ross 1930; Audi 2004.
13. Jackson 1986.
14. Enoch 2014.
15. Kauppinen 2017b.
16. Searle 1980.
17. Vrt. Himma 2009.
18. Ks. Visala 2018.
19. Fischer & Ravizza 1998.
20. Esim. Hew 2014.
21. Hew 2014; Hakli & Mäkelä 2016.
22. Mele 1995.
23. Vrt. Rodogno 2016.
24. Kauppinen 2017b.
25. Parfit 2011.
26. Audi 2004.
27. Esim. McDowell 1979.
28. Wallach & Allen 2008.
29. Wallach & Allen 2008.
30. Baars 1997.
31. Franklin 2003.
32. Chalmers 1996; Pylkkänen 2007.
33. Searle 1983.
34. Wallach & Allen 2008.
35. Picard 1995.
36. Torrance 2008; Sharkey 2017.
37. Floridi & Sanders 2004.
38. Coeckelbergh 2014; Gunkel 2018.
39. Coeckelbergh 2014, 71.
40. Dennett 1987.
41. Kuten tämän tekstin anonyymi arvioija huomautti, keinotekoiset oikeintekijät ovat yhdessä suhteessa takuuvarmempia kuin tavalliset moraalitoimijat, koska ne eivät kärsi tahdonheikkoudesta. Hyveelliset moraalitoimijat toki välttävät tämänkin ongelman, mutta on myönnettävä, että on kuviteltavissa tilanteita, joissa oikein ohjelmoitu

keinotekoinen toimija toimisi todennäköisemmin oikein kuin keskivertoihminen.
42. Esim. Darwall 2006.
43. Ks. Hakli & Mäkelä 2019.
44. Allen ym. 2000.
45. Karras ym. 2018.
46. Asimov 1950.
47. Arkin 2010.
48. Ks. Bykvist 2017.
49. Kauppinen, tulossa.
50. Aristoteles 2012, 100b21–33.
51. Esim. Ghandeharioun kumppaneineen (2017) on diagnosoinut masennusta ihon sähköaktiviteetin, unirytmien ja paikkatietojen avulla.
52. Smith 2002.

+ + +

Allen, Colin, Gary Varner & Jason Zinser (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12:3, 251–261.

Aristoteles (2012). *Topiikka/Sofistiset kumoamiset*. Suom. Juha Sihvola & Marke Ahonen. Helsinki: Gaudeamus.

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*. New York: Gnome Press.

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

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, e12408.

Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York & Oxford: 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 (2016). *Transit*. London: Jonathan Cape.

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

Dennett, Daniel C. (1987). *The Intentional Stance*. Cambridge, MA & London: MIT Press.

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

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

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

Floridi, Luciano & J. W. Sanders (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.
Ghandeharioun, Asma, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag & Rosalind Picard (2017). Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data. *Seventh International Conference on Affective Computing and Intelligent Interaction.* (ACII), San Antonio, TX, USA, 2017, 325–332. <doi.org/10.1109/ACII.2017.8273620>
Grice, Paul (1974–1975). Method in Philosophical Psychology (from the Banal to the Bizarre). *Proceedings and Addresses of the American Philosophical Association* 48, 23–53.
Grimm, Stephen R. (toim.) (2017). *Making Sense of the World: New Essays on the Philosophy of Understanding*. New York: Oxford University Press.
Gunkel, David J. (2018). The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology* 20:2, 87–99.
Hakli, Raul & Pekka Mäkelä (2016). Robots, Autonomy, and Responsibility. Teoksessa Johanna Seibt, Marco Nørskov & Søren Schack Andersen (toim.): *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016*. Amsterdam: IOS Press, 145–154.
— (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 (1983). *An Enquiry Concerning the Principles of Morals*. Toim. J. B. Schneewind. Indianapolis: Hackett (ilm. alun perin 1751).
Jackson, Frank (1986). What Mary Didn't Know. *Journal of Philosophy* 83 (May), 291–5.
Karras, Tero, Timo Aila, Samuli Laine & Jaakko Lehtinen (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *Proceedings of the International Conference on Learning Representations.* arXiv:1710.10196 [cs.NE]
Kauppinen, Antti (2015). Favoring. *Philosophical Studies* 172:7, 1953–1971.
— (2017a). Empathy and Moral Judgment. Teoksessa Heidi Maibom (toim.): *The Routledge Handbook of the Philosophy of Empathy*. London & New York: Routledge, 215–226.
— (2017b). Sentimentalism, Blameworthiness, and Wrongdoing. Teoksessa Karsten Stueber & Remy Debes (toim.): *Ethical Sentimentalism*. Cambridge: Cambridge University Press, 133–152.

— (tulossa). Who Should Bear the Risk When Self-Driving Vehicles Crash? *Journal of Applied Philosophy*.
McDowell, John (1979). Virtue and Reason. *The Monist* 62:3, 331–350.
Mele, Alfred R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. New York & Oxford: Oxford University Press.
Parfit, Derek (2011). *On What Matters: Two-Volume Set*. Oxford: 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*. Berlin & Heidelberg: Springer.
Rodogno, Rafael (2016). Robots and the Limits of Morality. Teoksessa M. Nørskov (toim.): *Social Robots: Boundaries, Potential, Challenges*. Farnham & Burlington: Ashgate, 39–55.
Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.
Schlosser, Markus E. (2012). Taking Something as a Reason for Action. *Philosophical Papers* 41:2, 267–304.
Searle, John (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3, 417–57.
— (1983). *Intentionality*. Cambridge: Cambridge University Press.
Sharkey, Amanda (2017). Can Robots Be Responsible Moral Agents? And Why Should We Care? *Connection Science* 29:3, 210–216.
Smith, Adam (2002). *A Theory of Moral Sentiments*. Toim. K. Haakonssen. Cambridge: Cambridge University Press (ilm. alun perin 1759).
Steiner, Adam P. & A. David Redish (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* 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*. Helsinki: Gaudeamus.
Wallach, Wendell & Colin Allen (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford & New York: Oxford University Press.
Westermarck, Edward (1906). *The Origin and Development of the Moral Ideas*. Freeport, N.Y.: Books for Libraries Press.