Michael Laakasuo
Jussi Palomäki

Robotiikan moraalipsykologian näkökulmia yhteiskuntaan ja työelämään

Tässä kirjoituksessa käsittelemme tutkimustemme merkitystä tulevaisuuden työelämälle. Tutkimusryhmämme tavoite on jalkauttaa ja vakinaistaa Suomeen uusi tutkimusala: robotiikan moraalipsykologia. Käymme ensin läpi moraalipsykologian nykytilaa ja esittelemme sen keskeisiä teorioita ja löydöksiä. Tämän jälkeen käsittelemme teknologisen nyky-yhteiskunnan luomia haasteita moraalipsykologialle, ja miten robotiikan moraalipsykologiaa voidaan soveltaa työelämässä. Keskeiset muutoksia työelämässä – moraalipsykologisesta näkökulmasta – tulevat olemaan, kohtaamiset laitteiden ja koneiden kanssa, jotka tekevät päätöksiä toisten ihmisten hyvinvoinnista. Tulevaisuuden työelämässä jokaiselta vaaditaan herkkyyttä ja ymmärrystä havaita älykkäät koneet moraalisesti relevantteina asioina, huolimatta siitä, ettei niillä ole varsinaista aitoa moraalitajua. Esittelemme eettisen sokeuden käsitteen ja selitämme, miksi ihmisen luontainen evoluution myötä muodostunut kognitio ei kykene ymmärtämään robotteja tai keinoälyjä oikealla tavalla. Projisoimme näihin laitteisiin helposti ominaisuuksia joita niissä ei ole, tai vaihtoehtoisesti olemme sokeita sille, että niihin pitää lähtökohtaisesti suhtautua kriittisesti. Pyrimme lukijaystävällisyyteen; emme oleta lukijalta aiempaa taustatietoa aiheesta.

AVAINSANAT: robotiikka, moraalipsykologia, tekoäly, työelämä

Kahdenlaista moraalipsykologia

Moraalipsykologia tarkastelee ihmisessä (ja muissakin eläimissä) niitä kognitiivisia eli tiedonkäsittelyn rakenteita, jotka liittyvät päätöksiin, arviointeihin ja neuvotteluihin oikeasta ja väärästä. Toisin kuin moraali*filosofia* (etiikka), moraalipsykologia tarkastelee moraalia empiirisenä, eli ihmisten (ja muiden eläinten) toiminnassa havaittavana ja mitattavana ilmiönä.

Moraali ilmenee ihmisten välisessä kanssakäymisessä. Moraalipsykologia harvemmin ottaa normatiivista kantaa asioihin tai pyrkii määrittelemään sitä, mikä on oikein tai väärin. Moraalifilosofi pyrkii selvittämään ennalta asetetuin eli *a priori* -menetelmin – kuten filosofisella käsiteanalyysillä ja filosofin intuitiolla – mahdollisia normatiivisesti sovellettavia ohjenuoria eri tilanteiden ymmärtämiseen. Moraalipsykologi puolestaan pyrkii havainnoimaan ja mittaamaan sekä tavallisten ihmisten että koulutettujen asiantuntijoiden intuitioiden, arvioiden, järkeilyjen ja päättelyiden ilmenemistä tilanteissa, jotka liittyvät tietoisten olentojen hyvinvointiin. Toistaiseksi moraalipsykologinen tutkimus on keskittynyt lähinnä ihmisiin, mutta viime aikoina myös eläinoikeuskysymykset (esim. Loughnan ym., 2010) ja robotiikan eettiset ongelmat (Bonnefon ym., 2016) ovat nousemassa moraalipsykologian huomion kohteeksi.

Moraalipsykologia jakautuu pitkälti kahteen melko erilliseen perinteeseen (Haidt, 2007). Moraalipsykologian ensimmäinen aalto (nk. kohlbergilainen moraalipsykologia) tarkasteli kehityspsykologisesta näkökulmasta moraalin kehityksen vaiheita, ja käytti hyväkseen syväluotaavia haastatteluaineistoja sekä laadullisia menetelmiä. Tässä koulukunnassa oletettiin, että moraalinen kyky ja ymmärrys kehittyvät ihmisille asteittain, edeten konkreettisesta rangaistuksen pelosta kohti kykyä soveltaa abstrakteja universaaleja moraalin periaatteita eri tilanteissa (esim. Helkama, 2009).

Oman tutkimusryhmämme menetelmät ja ajattelu sijoittuvat nk. moraalipsykologian toiseen aaltoon. Pyrimme luomaan ja löytämään moraalidilemmoja, tarinoita ja tilannekuvauksia, jotka asettavat vastakkain erilaisia moraali-intuitioita. Tämän lisäksi hyödynnämme ja sovellamme persoonallisuuspsykologian, käyttäytymistaloustieteen, evoluutiopsykologian ja neurotieteen tutkimusmenetelmiä ja teorioita. Moraalisen kognition tutkiminen, siten kuin ala ymmärretään moraalipsykologian toisen aallon näkökulmasta, on ”kovaa kvantitatiivista” laskennallista tiedettä, jossa tilastolliset menetelmät ovat pääroolissa ja laadulliset menetelmät tärkeässä sivuroolissa.

Nykyään moraalipsykologia nivoutuu läheisesti päätöksenteko- ja emootiotutkimukseen; monilla tieteenaloilla tehdään moraalisen kognition kannalta merkittävää tutkimusta, vaikka sitä ei aina kutsutakaan ”moraalipsykologiaksi”. Esimerkiksi kliiniset moraalipsykologit ovat kiinnostuneita psykopaattien ja narsistien tunne-elämästä, ja siitä, miten nämä ratkaisevat joitakin moraalisia ongelmia valtaväestöstä poiketen. Teologiaan perehtyneet moraalipsykologit puolestaan saattavat tarkastella, miten ihmisten uskomukset vapaasta tahdosta liittyvät avuliaisuuteen (Sinnott-Armstrong, 2017). Moraalipsykologia on toisin sanoen hyvin laaja kognition tutkimuksen ala, jota voidaan tarkastella monesta eri näkökulmasta, ja joka tuo yhteen niin kliinikoita, neurotutkijoita, kuin juristeja (esim. Mikhail, 2007) ja antropologejakin.

Moraalikognition keskeiset mallit

Moraalipsykologian toisen aallon tutkimusperinteeseen liittyy muutama keskeinen käsitteellinen malli. Näiden mallien ymmärtäminen on oleellista, jotta tutkimusryhmämme keskeisten teemojen tarkastelu voidaan asettaa oikeaan kontekstiin. Esittelemme malleja lyhyesti alla.

Minimimalli

Moraalikognition minimimallin kehittivät Kurt Grey kollegoineen (Gray ym., 2007; 2012). Heidän mukaansa moraalisen kognition mahdollistava keskeinen kognitiivinen mekanismi perustuu mielen projisointiin. Terveillä ihmisillä on automaattinen ja tiedostamaton taipumus havaita ”mielellistä kyvykkyyttä” toisissa elävissä olennoissa, etenkin toisissa ihmisissä – ihmiset siis projisoivat eli ”heijastavat” toisiin eläviin olentoihin samoja mielellisiä kykyjä, joita heillä itsellään on. Tämän mekanismin myötä ih-

miset projisoivat toisiinsa myös kyvyn tuntea kärsimystä, kyvyn motivoitua, tai kyvyn tehdä asioita tarkoituksenmukaisesti. Ihmiset havaitsevat tai tulkitsevat sosiaalisen tilanteen olevan moraalisesti merkitsevä, mikäli neljä ehtoa täyttyy: A1) toimija aiheuttaa A2) tarkoituksenmukaista vahinkoa B1) toiselle olennolle, joka B2) kykenee kokemaan kärsimystä.

Minimimallin kehittäjät pyrkivät mahdollisimman kattavaan selitykseen mahdollisimman yksinkertaisella mallilla. Keskeistä on se, että moraalinen kognitio, tai jonkin teon moraalisuuden arvioiminen, tehdään aina tilanteen ulkopuolelta, ”tarkkailijan” näkökulmasta. Malli pyrkii selittämään, mistä moraalisessa paheksunnassa on kysymys; malli ei kuitenkaan ota kantaa siihen, miksi joku teki vahinkoa toiselle, tai miten tekijän olisi kuulunut toimia.

Moraalitunteiden perusta

Sosiaalipsykologi Jonathan Haidtia (2012) pidetään yhtenä nykyisen moraalipsykologian uuden aallon keskeisenä kehittäjänä ja uudistajana. Haidtin kehittämän moraalipsykologisen mallin mukaan ihmisten moraalin perusta jakautuu viiteen osa-alueeseen (Graham ym., 2013). Arvioidessaan toisten tekojen moraalista hyväksyttävyyttä, ihmiset kiinnittävät huomiota siihen, 1) aiheutuiko teosta haittaa tai vahinkoa, 2) oliko teko reilu, 3) kunnioittiko teko auktoriteetteja, 4) oliko teko uskollinen sisäryhmälle, ja 5) oliko teko puhdas tai loukkasiko se pyhiä arvoja? Toisin kuin moraalikognition minimimallissa, Haidtin mallissa teon moraalinen paheksunta ei edellytä sitä, että teosta koituisi vahinkoa tai harmia tuntevalle oliolle. Haidtin (2012; Graham ym., 2013) malliin sisältyy myös oletus ihmisten välisistä yksilöllisistä eroista näissä viidessä moraalin osa-alueessa. Hänen mukaansa – ainakin Yhdysvalloissa, joskaan ei yhtä selkeästi Suomessa tai muualla Euroopassa – liberaalisti suuntautuneille ihmisille on tekojen moraalia arvioitaessa merkitystä erityisesti sillä, oliko teko reilu ja koituiko siitä vahinkoa. Konservatiiviset ihmiset puolestaan painottavat moraaliarvioissaan liberaaleja enemmän myös auktoriteetin kunnioittamista, uskollisuutta sisäryhmää kohtaan ja teon ”puhtautta”. Konservatiivit paheksuvat maan lipun polttamista tai kannabiksen käyttöä enemmän kuin liberaalit. Liberaalit taas todennäköisemmin paheksuvat tuloeroja ja niiden tarkoituksenmukaista kasvattamista.

Moraaliset emootiot ohjaavat käyttäytymistä

Moraalisen kognition tutkimus on kiisteltyä. Eräs kiistelyn kohde liittyy hieman yksinkertaistettuna siihen, perustuuko moraalinen päätöksenteko tunteisiin vai järkeen (esim. Greene, 2013). Jonathan Haidt (2007) on aikaisemmin esittänyt, että moraaliarviot perustuvat pitkälti tunnereaktioihin: tuomitsemme tekoja siksi, että ne herättävät negatiivisia emootioita kuten inhoa tai suuttumusta. Nykyään kuitenkin ajatellaan, että myös tietoisesti valituilla periaatteilla, pohdinnalla, tai yleisemmin ”moraalisella järjellä” on merkittävä rooli tekojen moraalisessa arvioinnissa (McAuliff, arviossa; Mikhail, 2007).

Emootioiden rooli moraaliarvioissa korostuu tilanteissa, joissa painotetaan yleisten sääntöjen tai normien noudattamista. Mikäli henkilö esimerkiksi uskoo, että tappaminen on aina väärin, tappamiseen liittyvien emootioiden, kuten suuttumuksen tai inhon, merkitys korostuu tekoa tuomittaessa. Asia ei kuitenkaan ole aivan näin yksinkertainen. Seurauseettisestä eli utilitaristisesta näkökulmasta katsoen tappaminen voi olla oikeutettua, jos yhden ihmisen tappaminen johtaa useamman pelastumiseen (Greene, 2007; 2013). On ajateltu, että utilitaristinen ”moraalilaskenta” on vähemmän emotionaalista kuin absoluuttisiin sääntöihin, normeihin ja yleisiin velvollisuuksiin perustuva moraali (nk. velvollisuuseettinen eli deontologinen moraali; Greene, 2013).

Utilitaristisen moraalikäsityksen mukaan tarkoituksenmukaista on pyrkiä maksimoimaan ”hyvän” tai hyvinvoinnin määrä tilanteesta riippumatta. Utilitaristinen moraali asetetaan usein vastakkain deontologisen moraalin kanssa, ikään kuin moraalinen toimija olisi lähtökohtaisesti joko utilitaristinen tai deontologinen. Aivokuvantamistutkimuksissa on aiemmin havaittu, että deontologiset moraaliarviot tehtiin utilitaristisia arvioita nopeammin, intuitiivisemmin, ja niihin liittyi aktivaatiota ”primitiivisimmillä” aivojen alueilla (esim. Greene, 2013). Vastaavasti utilitaristiset moraaliarviot näyttivät alussa korreloivan ”korkeampaan kognitioon” (työmuisti, rationaalinen toiminnan ohjaus, reflektointi) liittyvien aivojen alueiden aktivaation kanssa. Tästä pääteltiin, että utilitaristinen moraali on deontologiseen moraaliin verrattuna sekä kognitiivisesti raskaampaa että reflektiivisempää, ja myös vähemmän tunnepohjaista ja intuitiivista. Tuoreempi tutkimustieto on toisaalta kyseenalaistanut aiempia tuloksia; esimerkiksi aivovauriot (Koenigs ym., 2007; Christensen & Gomila, 2012), psykopatia (Koenigs ym., 2011; Bartels ym., 2011), kyvyttömyys kokea tunteita (Patil ym, 2014) ja erilaiset akuutit päihtymystilat saattavat lisätä ihmisten taipumusta utilitaristisiin moraaliarvioihin (Duke & Bègue, 2015; Perkins ym., 2013). Lisäksi on lähtökohtaisesti ongelmallista väittää, että hidas, aktiivista päättelyä ja työmuistitoimintoja vaativa utilitaristinen moraalikognitio olisi jollain tapaa vähemmän alkeellista (tai ”parempaa”) kuin deontologinen moraalikognitio. Oikeastaan nopeat emootiot, intuitiot ja muut tunteet ovat yhtä lailla kognitiota (eli informaation prosessointia) kuin työläs asioiden järkeily. Automaattinen kognitio saattaa hyvinkin prosessoida ja integroida suuremman määrän informaatiota, kuin tiedostetumpi ja ”hitaampi” kognitio. Suurin osa ihmisen toiminnasta perustuu kuitenkin massiiviselle määrälle tietoisuuden ulkopuolella tapahtuvaa kognitiivista prosessointia.

Emootioista inho, suuttumus ja halveksunta näyttävät tavalla tai toisella liittyvän moraaliarvioihin. Tiedetään, että suuttumus tai moraalinen raivo motivoi rankaisuun: jos joku loukkaa meitä, varastaa meiltä, tai kohtelee meitä muuten kaltoin, suutumme, ja motivoidumme rankaisemaan tätä henkilöä. Tämä puolestaan viestii meitä vastaan rikkoneelle henkilölle, että hänen toimintansa ei ollut kannattavaa. Myös inholla on merkittävä, mutta kiistanalainen rooli moraalikäyttäytymisessä (esim. Laakasuo ym., 2017; Tybur ym., 2013). Tyypillisesti on ajateltu, että inhon tunne toimii yhtenä moraalisen paheksunnan portinvartijana; inhottavat asiat laukaisevat paheksunnan. Inho on kuitenkin monimutkainen emootio, joka voi liittyä esimerkiksi patogeeneihin (bakteerit), abstrakteihin asioihin (esim. lipun polttaminen) tai seksuaalisuuteen (”epäsopivat” kumppanit tai valtavirrasta poikkeava seksuaalisuus). Inhon eri muotoja voidaan selittää evoluutioteorian avulla, mutta emme paneudu näihin selityksiin tässä esseessä (ks. Tybur ym., 2013).

Olemme omissa tutkimuksissamme tarkastelleet ihmisten taipumuksia kokea inhoa, eli ns. inhoherkkyyttä. Inhoherkkyys ei ole sama asia kuin jossain tietyssä tilanteessa koettu inho. Inhon kokeminen sinänsä ei liity vahvasti moraaliin, mutta herkkyys kokea inhoa liittyy. Toisin sanoen, mitä herkemmin ihminen reagoi inhottavina pitämiinsä asioihin, sitä vahvemmin hän paheksuu eri asioita, jopa sellaisia, jotka eivät päällisin puolin liity inhottaviin asioihin. Patogeeneihin liittyvä inhoherkkyys näyttää ennustavan erityisesti utilitaristista moraalia, kun taas seksuaali-inhoherkkyys deontologista moraalia (Laakasuo ym., 2017). Sekä utilitaristiset että deontologiset moraaliarviot liittyvät tunnejärjestelmiin; toisaalta ihmisen moraalikognition ja tunnejärjestelmien välinen yhteys on yhä osittainen mysteeri.

Evoluutiopsykologia ja uusi ontologinen kategoria

Evoluutiopsykologisessa kognition tutkimuksessa ja kognitiivisessa antropologiassa on tar-

kasteltu ns. luonnollisia kategoriota tai intuitiivista biologiaa (esim. Atran, 2012; Boyer & Barrett, 2015). Näillä tarkoitetaan ihmisten synnynnäisiä valmiuksia luokitella ympäristön ärsykkeitä automaattisesti ja refleksinomaisesti. Pienet lapset osaavat intuitiivisesti (kykenemättä välttämättä pukemaan toimintaansa sanoiksi) luokittelemaan petoeläimet ja saaliseläimet omiin lokeroihinsa; ja vastaavasti kasvit eläimistä tai esimerkiksi kivistä erillisiin lokeroihin. Tyypillisesti tarhaikäisten lasten leikeissä seeprat eivät syö leijonia, eivätkä puut kävele ja syö seeproja (Boyer & Barrett, 2015). Koira ei myöskään tämän ikäisten lasten mielestä oikeasti muutu norsuksi, vaikka sille liimattaisiin kärsä ja isot korvat (Gelman & Wellman, 1991). Vastaavasti esipuheikäiset lapset hämmästyvät, jos näennäisesti kiinteä esine kulkee toisen kiinteän esineen läpi pysähtymättä tai törmäämättä siihen (Moll & Tomasello. 2010). Ihmisellä on toisin sanoen hyvin nuoresta iästä lähtien luontaisia käsityksiä kategorioista, kuten ”eläimenä oleminen”, ”kiinteys” tai ”olemuksellinen pysyvyys”.

Luontainen tai synnynnäinen intuitiivinen ymmärrys maailmasta mukailee oikean maailman rakennetta; ei täydellisesti, mutta niin hyvin, että se on estänyt meitä evoluution aikana tekemästä vakavia virhearviointeja selviytymisemme kustannuksella. Ihmisillä näyttäisi olevan jopa luontaisia kognitiivisia mekanismeja työkalujen luokittelemiseksi ja tunnistamiseksi: aivomme aktivoituvat eri tavalla, jos katsomme videoita vierinkivikulttuurin (1.5–2.5 miljoonaa vuotta sitten) kivityökalujen valmistamisesta, verrattuna acheulin kulttuurin (200 tuhatta–1.2 miljoonaa vuotta sitten) aikaisten kivityökalujen valmistamisen katseluun. Acheulin kulttuuri on evolutiivisessa aikaskaalassa mitattuna selkeästi lähempänä oman lajimme syntyä ja kehitystä (ks. Putt ym., 2017).

Olemme evoluutiohistoriamme aikana kehittyneet ympäristössä, jossa olemme luontaisesti havainneet muita ihmisiä (lapsia, aikuisia), petoeläimiä, työkaluja, ja esimerkiksi ”kiinteyttä”. Meillä on luontainen valmius ymmärtää toistemme tunteita ja aikeita, kyky varautua petoeläimistä ja muista vaaroista aiheutuviin uhkiin, ja halu auttaa ketä tahansa vaaraan joutunutta lasta. Robotit ja tekoäly eivät kuitenkaan ole luontaisia kategorioita evoluution muokkaamille aivoillemme ja kognitiollemme (Severson & Carlson, 2010). Ne aktivoivat aivojamme täysin uusilla ja hankalasti ennakoitavilla tavoilla. Olemme luontaisesti eettisesti sokeita (Palazzo ym., 2012) robottien ja tekoälyjen muodostamille riskeille ja mahdollisuuksille (Tegmark, 2017; Bostrom, 2014).

Nykyaikaisten tekoälyjen ohjelmoiminen ja rakentaminen, ja niiden yksityiskohtainen ymmärtäminen, edellyttävät korkeatasoista ja hienostunutta matematiikkaa – suurimmalle osalle meistä monimutkaisten tekoälyjen ohjelmointi ja toiminta on intuitiivisesti mahdotonta ymmärtää. Tekoälyjen toiminnan tarkka ja kypsä ymmärtäminen niiden mahdollisuuksien ja rajoitteiden puitteissa on verrattavissa mihin tahansa kulttuurisesti monimutkaiseen taitoon, jonka omaksumiseen menee vuosia (Tegmark, 2017; Bostrom, 2014). Kehityspsykologit ja robotiikan filosofian tutkijat ovatkin ehdottaneet, että tekoälyt ja robotit pitäisi nähdä uutena ontologisena kategoriana (Severson & Carlson, 2010): ne eivät ole eläviä, mutta eivät varsinaisesti kuolleitakaan (ne liikkuvat, tai vähintäänkin tekevät asioita itsenäisesti). Ne ovat olemassaolon muotona jotain täysin uutta maapallon historiassa.

Robotiikan moraalipsykologia työelämässä

Robotiikka ja tekoälyt yleistyvät työelämässä ja niistä on viime vuosien aikana käyty aktiivista keskustelua. Robotiikan moraalipsykologian osuus tästä keskustelusta on valitettavasti jäänyt hieman ontoksi; suurin osa siitä on ollut Suomessa meidän tutkimusryhmämme vastuulla. Robotiikan moraalipsykologialla on paljon an-

nettavaa kaikille organisaatioille, jotka suunnittelevat tietohallintokäytänteitä, laitehankintoja, ja haluavat ymmärtää miten ihmiset reagoivat tulevaisuuden teknologioihin sekä tunteiden että käyttäytymisen tasolla. Tulevaisuudessa käytössämme on älykkäitä proteeseja, kuten aivoihin asennettavia muisti-implantteja. Tehokkaat profilointi-algoritmit priorisoivat potilaita leikkausjonoihin, ja hoitorobotit tekevät osittain itsenäisiä hoitopäätöksiä. Liikenteessä itseohjautuvien autojen määrä kasvaa räjähdysmäisesti, ja joudumme tekemisiin automaattipoliisien, -juristien ja -kirjanpitäjien kanssa. Älykäs teknologia alkaa pian tehdä (ja tekee jo nyt) ihmisen hyvinvointiin liittyviä päätöksiä, ja tämä on moraalipsykologinen ongelma. Minkälaiset algoritmiset ratkaisut herättävät paheksuntaa esimerkiksi potilasjonojen priorisoinnissa tai potilaiden lääkitsemisessä? Miten tulisi suhtautua itseohjautuvaan autoon, joka ajaa lapsen yli? Miten ihmisiä pitäisi kouluttaa ja valistaa, jotta he havaitsevat eettisesti sokeat pisteet (Palazzo ym., 2012) tekoälykehityksessä ja teknologiassa? Mitkä moraalituntojen perustoista ovat tärkeitä tälle kokonaisuudelle ja sen hahmottamiselle, vai ovatko mitkään?

Aivoimplantit, seksuaali-inho ja työelämän reiluus

Tulevaisuuden ennustaminen on ihmiselle erittäin vaikeaa; parhaat poliittiset ennustajat kykenevät ennustamaan poliittisia tapahtumia vain noin 300 päivän päähän, ja ennustaminen ei ole tyypillisesti kovin tarkkaa (Mellers ym., 2015). Emme tässä hetkessä kykene ennustamaan tai kuvailemaan sitä, miltä tekoäly näyttää tai mihin se pystyy kymmenen vuoden, saati sitten 50 vuoden kuluttua (Tegmark, 2017; Bostrom, 2014). Olemme toisin sanoen mielikuvituksellisesti sokeita älykkään teknologian todellisille mahdollisuuksille tulevaisuudessa. Teknologian alalla työskentelevillä on suuri vastuu suunnitella tekoälyjä, sillä on mahdollista, että jo 50 vuoden kuluttua tekoäly on ohittanut ”ihmisälyn” kaikilla sen osa-alueilla (Tegmark, 2017). Se, mitä ihmistä älykkäämpi, ihmisen luoma keinotekoinen äly pystyy tai tulee tekemään, on yhtä vaikeaa ennustaa kuin mustaan aukkoon katsominen – tästä johtuen tekoälyn ”älykkyysräjähdystä” kutsutaan myös teknologiseksi singulariteetiksi (Tegmark, 2017; Bostrom, 2014).

Muita tulevaisuudessa älykkään teknologian kehittymiseen liittyviä kysymyksiä ovat mm. aivoihin asennettavien mikrosirujen käyttöön liittyvät ongelmat. Aivoihin asennettavat mikrosirut eivät ole vielä aktiivisessa käytössä työelämässä, mutta asia voi hyvinkin olla toisin jopa 20 vuoden sisällä. Pitäisikö tulevaisuuden työelämässä olla mahdollista, tai jopa velvollisuus, asentaa aivoihin mikrosiruja, jotka mahdollistavat suoran kommunikaation Internetin välityksellä muiden ihmisten (aivojen) kanssa? Työelämän suoritusvaatimusten kasvaessa ei ole selvää, kuinka helppo yksilöiden on kieltäytyä erilaisten teknologioiden käyttöönotosta. Omissa työn alla olevissa tutkimuksissamme olemme havainneet, että ihmiset, joilla on korkea seksuaalinen inhoherkkyys eivät pidä siitä, että toiset ihmiset asentavat aivoihinsa mikrosiruja; esimerkiksi korjatakseen alkavia muistisairauksia tai palauttaakseen muistinsa toiminnan nuoruusiän tasolle. Lisäksi näiden ”aivoproteesien” käytön hyväksyttävyys laskee dramaattisesti, mikäli mikrosiruilla on mahdollista saada yli-inhimillisiä kykyjä. Aiemmissa tutkimuksissa on havaittu, että teknologian käyttöä (esim. doping) paheksutaan, jos sen koetaan muuttavan ”ihmisyyden ydintä”, tai jos ihmisen inhimillisyys esineellistetään (Sthrominger & Nichols, 2014).

Profilointialgoritmit ja lainvalvonta

Tulevaisuudessa myös poliisityö saattaa muuttua merkittävästi (ks. O’Neil, 2017). Algoritmeihin perustuva lainvalvonta ja ihmisten profilointi tulevat yleistymään ja vaikuttamaan yhteiskunnan yleisilmapiiriin. Jos tulevaisuudessa yhteiskunnassa vallitsee riittävän pelokas (esimerkiksi sodista johtuva) ilmapiiri, valtio voi hyödyntää tehokkaita profilointialgoritme-

ja, joilla on pääsy kaikkiin mahdollisiin rekistereihin. Kerätyn tiedon perusteella voidaan muodostaa automaattinen ”uhka-arvio” jokaisesta kansalaisesta, ja poliisi voi mukauttaa toimintaansa näihin arvioihin perustuen – ilman, että yksikään ihmispoliisi on edes tietoinen algoritmin toimintaperusteista.

Kuvitellaan Marko, joka menee parturiin ja rastoittaa hiuksensa. Tämän jälkeen hän matkustaa Amsterdamiin, ja kotiin palattuaan käy ”psy-trance” musiikkitapahtumassa. Markon Spotify -soittolistalla on King Crimsonia, Toolia, Pink Floydia, ja Shponglea. Markon luottokortilla näkyy ostotapahtumina Boom-festivaalin lippu, luomuruokaa, altakasteluruukkuja, kasvilamppuja, soraa, ja kasvilannoitteita. Psykologisen tutkimuksen valossa tiedämme jo nyt, että kyseiset musiikkivalinnat, matkustuskohteet, ajanviettotavat ja hiustyylit liittyvät päihdemyönteisiin asenteisiin. Ei olisi siis ihme, jos tulevaisuudessa poliisi saisi automaattiselta profilointialgoritmilta hälytyksen käydä pidättämässä Marko mahdollisen huumerikoksen valmistelusta. Tätä kirjoittaessa maailmalla työstetään jo juridiikkapalveluiden, kuten kirjanpidon ja perunselvitysten automatisointia; pidätysluvan prosessoiminen riittävän epäilyn perusteella voitaisiin täysin realistisesti myös automatisoida. Tämän jälkeen poliisi voisi käydä pidättämässä Markon rikoksen valmistelusta ”riittävin” perustein.

Tulevaisuuden profilointialgoritmit eivät voi välttyä nk. vääriltä positiivisilta. On täysin mahdollista, että Marko olisi saattanut haluta vain kasvattaa chilipaprikoita, koska siihen tarvittavat välineet sattuvat olemaan samoja kuin kannabiksen viljelyyn tarvittavat välineet. Yhteiskunta, joka torjuu ennalta kaikki rikokset, on tietystä näkökulmasta turvallinen; mutta siihen liittyy myös merkittäviä ongelmia (samaa aihetta on käsitelty mm. Tom Cruisen tähdittämässä vuoden 2002 skifi-toimintaelokuvassa ”Minority Report”). Kuinka suuresta osasta yksityisyyttämme olisimme valmiita luopumaan, jotta yhteiskunta olisi ”täydellisen” turvallinen? Kysymys on toistaiseksi teoreettinen, mutta ei välttämättä enää kauan.

Yllä kuvattu ongelma on jo nyt moraalipsykologian kannalta merkittävä. Emme vielä tiedä, tai oikeastaan edes kattavasti tutki, kuinka moni ihminen pitäisi esitetyn kaltaista ”täydellisen turvallista” kansalaisia valvovaa ja profiloivaa yhteiskuntaa toivottavana. Voi hyvin olla, että konservatiivit ovat taipuvaisempia koko moraalituntojen paletilla kannattamaan tällaisia teknologioita ja niiden käyttöönottoa. Tiedämme myös, että suurempi osa kansasta on konservatiiveja kuin liberaaleja. Tulevaisuuden uhkakuvana voi olla myös, että ihmiset moraalisiin tuntoihinsa ja intuitioihinsa luottamalla, kannattavat sellaisten järjestelmien käyttöönottoa, joka tekee länsimaista enemmän Kiinan ja Turkin kaltaisia totalitaristisia valtioita. Moraalipsykologia ei ota kantaa siihen, onko tämä hyvä vai huono asia. Olisi kuitenkin tärkeää käydä keskustelua siitä, että halutaanko näihin intuitioihin nojaamalla luoda rakenteita, jotka vaarantavat demokratian perustoimintaedellytykset ja peruskansalaisvapauksien toteutumisen. On mahdollista, että moraaliset intuitiomme, jotka demokratiassa saavat toimia kohtuullisen vapaasti, uuden teknologian myötä tuhoavat sen pohjan, jonka varaan tämä vapaus on alunperin rakentunut. Tämä muuttaisi myös työelämän miellyttävyyttä, velvoittavuutta ja ahtautta.

Omien tutkimustemme tulokset viittaavat siihen, että ihmisillä ilmenee tähän liittyen myös utilitaristinen vääristymä: jos uusi teknologia vaikuttaa turvalliselta ja lisäävän merkittävästi hyvinvointia yhteiskunnassa, ihmiset ovat valmiita hyväksymään sen käytön. Ihmiset eivät kuitenkaan osaa helposti kuvitella itseään tulevaisuuden teknologian käytön tai soveltamisen uhreina. Ihmisillä ei ole luontaista kykyä tai tapaa nähdä näitä uhkia ja niiden mahdollista toteutumisia tarkasti ja monipuolisesti.

Monilla Internet -foorumeilla on tuoreeltaan

keskusteltu viimeaikaisista onnettomuuksista, jotka koskivat itseohjautuvia autoja. Itseohjautuvien autojen käyttöä puolustetaan tyypillisesti väittämällä niiden säästävän pitkällä aikavälillä ihmishenkiä. Ihmiset suosivat utilitaristisia moraalipäätöksiä, paitsi jos he joutuisivat itse olemaan ”muiden puolesta uhrattavan” roolissa. Haluaisimme liikenteeseen sellaisia itseohjautuvia autoja, jotka pyrkivät vaaratilanteessa aina minimoimaan uhrien määrän – mutta emme haluaisi olla tällaisten autojen kyydissä, koska ne eivät aina välttämättä suojele kuljettajaansa.

Vakuutusyhtiöt käyttävät jo nyt profilointialgoritmeja. Tulevaisuudessa henkilön googlailuhistoria saattaa johtaa kielteiseen vakuutuspäätökseen, mikäli kyseinen henkilö on viime aikoina esimerkiksi etsinyt usein tietoa jostain sairaudesta – riippumatta siitä, onko hän itse sairas vai ei. Profilointialgoritmeja voidaan myös käyttää työntekijöiden rekrytoimisen tukena, jolloin työhaastattelun ja ihmisten kesken käydyn neuvottelun merkitys työnhaussa vähenee merkittävästi. Moraalipsykologisesti näitä tilanteita voidaan tutkia esimerkiksi selvittämällä, miten epäoikeudenmukainen profilointialgoritmin päätöksen pitäisi olla, ennen kuin sen käyttöä aletaan pitää ongelmallisena. Miten deontologiset moraali-intuitiomme toimivat tilanteessa, jossa algoritmit kohtelevat meitä välineinä, joilla yritys voi maksimoida tuottonsa? Vai olemmeko me tälle asialle eettisesti sokeita?

On myös mahdollista, että profilointialgoritmien käyttö nähdään riskeistään huolimatta hyväksyttävänä; esimerkiksi koska ne vapauttavat ihmiset vaikeista päätöksistä ja siten myös vastuusta. Amerikkalaisen United Airlines lentoyhtiön lennolta jouduttiin 2017 poistamaan väkivalloin matkustaja, joka ei suostunut luopumaan ylivaratusta paikastaan. Koska vapaaehtoisia lennolta lähtijöitä ei löytynyt, tietokonealgoritmi päätti, kenet poistetaan koneesta. Itse tilanteessa kukaan lentohenkilökunnan jäsen ei kyseenalaistanut järjestelmän päätöstä, eikä suostunut joustamaan lainkaan. Tilanne päättyi väkivaltaiseen vastahakoisen matkustajan poistamiseen koneesta, jonka seurauksena tämä sai aivotärähdyksen ja menetti hampaan. United Airlinesin osake laski satoja miljoonia tapauksen jälkeen, ja yhtiö maksoi lopulta matkustajalle merkittävän korvauksen. Algoritmien tekemät päätökset eivät johda työntekijöiden irtisanomisiin. Jos United Airlinesin työntekijät olisivat säännöistä poiketen tarjonneet esimerkiksi isomman rahasumman korvaukseksi lennolta lähtemiseen, olisi tilanteelta varmasti vältytty. On kuitenkin helpompi toimia sokeasti sääntöjen mukaan, ja luottaa tietokonejärjestelmiin, koska tällöin välttyy varmasti ongelmilta esimiestensä kanssa. Tässä tilanteessa on myös läsnä sekä eettistä sokeutta että uuden ”ontologisen kategorian” muodostama havainto- tai ajatteluvirhe.

Terveyspalveluiden automatisoituminen ja moraalipsykologia

Moraalipsykologisella tutkimuksella on paljon annettavaa terveydenhuoltopalveluille. Tätä kirjoitettaessa pohditaan jo keinoja ottaa IBM:n Watsonin kaltaisia tekoälypalveluita diagnostiikan apuvälineiksi. Terveydenhoidon hallinnoinnin päätöksenteko on melko läpinäkymätöntä; resurssien priorisoinnissa ja kohdentamisessa on paljon parantamisen varaa. Monen arkijärkinen oikeustaju sanoo, että mikäli uutta maksaa itselleen odottavat elinsiirtojonossa 10-vuotias reipas koululainen ja 60-vuotias työtön alkoholisti, niin lapsi ”ansaitsisi” maksan 60-vuotiasta enemmän. Lain mukaan potilaita ei kuitenkaan saa asettaa eriarvoiseen asemaan. Käytännössä hallinnossa tehdään kuitenkin lukuisia ”näkymättömiä” päätöksiä ja tarjotaan kafkamaisen kimurantteja selityksiä sille, miksi leikkausjonon järjestystä pitää muuttaa.

Tiedämme henkilökohtaisten yrityskontaktiemme kautta, että Suomessa on yrityksiä, jotka haluavat tehdä leikkausjonojen priorisointipäätöksistä läpinäkyviä tekoälyteknologiaa hyödyntämällä. Hallinnollisten päätösten

automatisointi ja leikkausjonojen priorisointialgoritmin tulisi kuitenkin olla julkista tietoa ja kaikkien nähtävillä. Tällä hetkellä ei ole olemassa selkeää ohjeistusta siihen, millä perusteella lääkäri voisi muuttaa leikkausjonojen järjestystä. Tästä syystä aihetta olisi syytä tutkia moraalipsykologian näkökulmasta. On myös mahdollista, että tulevaisuudessa valtioiden velkaantumisen ja taloudellisen paineen johdosta sosiaali- ja terveyspalveluita halutaan automatisoida mahdollisimman tehokkaasti. Ei ole mitenkään mahdotonta kuvitella tilannetta, jossa tulevaisuudessa IBM:n Watsonin tyyppiset tekoälyt joutuvat tekemään diagnostisia päätöksiä ja antamaan hoito-ohjeita hoitoroboteille tai muille automaateille, jotka annostelevat lääkkeet vuodepotilaalle – mahdollisesti jopa näitä näkemättä tai kuulematta. Tälläkin hetkellä ihmisistä voidaan kerätä suuria määriä biometristä tietoa, jos näin halutaan tehdä. Sydämen sykkeen, verenpaineen, veriarvojen, ihon sähkönjohtavuuden ja monien muiden fysiologisten parametrien reaaliaikainen seuraaminen ja tallentaminen on jo nyt täysin mahdollista. Jos automatisoitu hoitodiagnostiikka tai sitä seuraavat hoitopäätökset sisältävät virheitä, on hyvin haastavaa sanoa, kuka tällöin on vastuussa virheistä. Watsonin kehittäjä ei itsekään omien sanojensa mukaan tiedä tai ymmärrä, miten tekoäly päätyy niihin tuloksiin, joihin se päätyy (Barratt, 2011).

Olemme omissa kokeissamme tarkastelleet tilanteita, joissa potilas kieltäytyy ottamasta lääkkeitään vastoin ylilääkärin antamia hoitomääräyksiä. Hoidosta vastaava sairaanhoitaja on joko ihminen tai robotti, joka joko noudattaa potilaan tahtoa ja jättää lääkkeet antamatta tai noudattaa lääkärin määräyksiä ja pakkolääkitsee potilaan tämän tahdon vastaisesti. Yleisesti ottaen pakkolääkitsemistä paheksutaan, jos sen tekee hoitorobotti, mutta hyväksytään, jos sen tekee ihmishoitaja. Hoitopäätöksen arvioon vaikuttaa myös se, kuoleeko potilas päätöksen seurauksena – erityisesti ihmishoitajan toimintaa arvioidessa. Myöskään hoitorobotin toimintavarmuudella (luotettava robotti vs. epävarmasti toimiva robotti) ei ole yhteyttä siihen, miten hyväksyttävänä sen toimintaa pidetään. Vastaavasti ihmishoitajan maineella on suuri merkitys hänen toimintansa arvioimisessa: huonomaineisen ihmishoitajan päätös olla noudattamatta ylilääkärin ohjeita koetaan hyvin paheksuttavaksi. Saatamme toisin sanoen eettisen sokeuden johdosta asettaa ihmisille täysin erilaisia moraalisia kriteereitä kuin koneille; ja lisäksi saatamme tehdä epäoptimaalisia laitehankintoja, jos emme ymmärrä vaatia koneilta erinomaista suoritustasoa.

Terveyspalveluissa passiivinen eutanasia on kiistelty aihe, jossa voitaisiin niin ikään soveltaa algoritmiikkaa. Missä tilanteessa elämää ylläpitävät laitteet saisi laittaa pois päältä; ja onko ihmisen tunteiden ja oikeustajun suhteen merkitystä, tekeekö päätöksen algoritmi vai ihminen (tai ihminen algoritmin avustuksella)? Näihin merkittäviin kysymyksiin ei ole olemassa vielä vastauksia, koska aihetta ei ole vielä tutkita kunnolla. Olemme kuitenkin aloittaneet tämänkin kysymyksen tarkastelun ja ensimmäiset kokeet ovat valmisteilla. Selvää kuitenkin on, että tämä teknologia tulee tavalla tai toisella muuttamaan sote-sektorilla toimivien ihmisten arkea.

Lopuksi

Olemme kirjoituksessamme pohtineet robotiikan moraalipsykologian merkitystä nyky-yhteiskunnassa ja erityisesti työelämän eri osa-alueilla. Koneiden moraalin tutkimus on vasta aluillaan. Emme ole kirjoituksessamme pohtineet kattavasti esimerkiksi itseohjautuviin autoihin liittyviä moraalipsykologisia ongelmia; tätä tutkitaan parhaillaan MIT:ssä suurella budjetilla. Keskityimme käsittelemään sitä, minkälaisia haasteita tulevaisuuden teknologiat voivat asettaa sekä yksittäisille ammattiryhmille että yleisesti sille, minkälaisessa yhteiskunnassa haluaisimme elää. Olemme keskittyneet ihmisten tunteisiin ja ajatuksiin siitä, mitä oikeaan ja

väärään liittyvät kokemukset merkitsevät teknologian keskellä uudistuvassa työelämässä. Olemme myös korostaneet, että ihmisen evoluutiohistorian myötä kehittynyt kognitiomme ei kykene luontaisesti jäsentämään intuition tasolla nyky-ympäristön teknologiailmiöitä – robotit ja muut tekoälyt ovat meille hyvin ”epäjatkuvuuspillisiä” asioita. Esimerkiksi robottikoiran potkaiseminen tuntuu meistä usein ”väärältä”, vaikka sillä ei ole tunteita, kokemuksia tai suunnitelmallista ajattelua. Olemme lihaan ja vereen savanniapinoita, jotka elävät mekaanisten zombien kanssa keskellä metalliviidakkoa. ■

LÄHTEET

Atran, S. (2012). Psychological origins and cultural evolution of religion. Grounding Social Sciences in Cognitive Sciences, 209-238.

Barrat, J. (2011). Our Final Invention. New York: Thomas Dunne.

Bartels, D. M. & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. Cognition, 121, 154-161.

Bonnefon, J. F., Shariff, A. & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352, 1573-1576.

Bostrom, N. (2014). Superintelligence. Oxford: Oxford University Press.

Boyer, P. & Barrett, H. C. (2015). Intuitive Ontologies and Domain Specificity. The Handbook of Evolutionary Psychology. New Jersey: Wiley & Sons Inc.

Christensen, J. F. & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. Neuroscience & Biobehavioral Reviews, 36, 1249-1264.

Duke, A. A. & Bègue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. Cognition, 134, 121-127.

Gelman, S. A. & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. Cognition, 38, 213-244.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P. & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In Advances in experimental social psychology (Vol. 47, pp. 55-130). Academic Press.

Gray, H. M., Gray, K. & Wegner, D. M. (2007). Dimensions of mind perception. Science, 315, 619-- 619.

Gray, K., Young, L. & Waytz, A. (2012). Mind perception is the essence of morality. Psychological Inquiry, 23, 101 – 124.

Greene, J. D. (2007). The secret joke of Kant’s soul. Teoksessa W. Sinnott-Armstrong (ed.), Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Disease, and Development (p. xx-yy). Cambridge, MA: MIT Press.

Greene, J. D. (2013). Moral Tribes. London: Penguin

Haidt, J. (2012). The Righteous Mind. London: Penguin

Haidt, J. (2007). The new synthesis in moral psychology. Science, 316, 998-1002.

Helkama, K. (2009). Moraalipsykologia. Helsinki: Edita.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M. & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. Nature, 446, 908.

Koenigs, M., Kruepke, M., Zeier, J. & Newman, J. P. (2011). Utilitarian moral judgment in psychopathy. Social Cognitive and Affective Neuroscience, 7, 708-714.

Laakasuo, M., Sundvall, J. & Drosinou, M. (2017). Individual differences in moral disgust do not predict utilitarian judgments, sexual and pathogen disgust do. Scientific reports, 7, 45526.

Loughnan, S., Haslam, N. & Bastian, B. (2010). The role of meat consumption in the denial of moral status and mind to meat animals. Appetite, 55, 156-159.

McAuliff, W. (arviossa). Do emotions play an essential role in moral judgments? https://psyarxiv.com/ajbc9/ Noudettu 16.04.2018

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M. & Ungar, L. (2015). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science, 10, 267-281.

Moll, H. & Tomasello, M. (2010). Infant cognition. Current Biology, 20, R872-R875.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. Trends in Cognitive Sciences, 11, 143-152.

Russell S. & Norwig P. (2015). Artificial Intelligence: A Modern Approach. London: Pearson Education

Strohminger, N. & Nichols, S. (2014). The essential moral self. Cognition, 131, 159-171.

Tegmark, M. (2017). Life 3.0. New York: Knopf

Wallach, W. & Allen. C. (2009). Moral Machines. Oxford: Oxford University Press

MICHAEL LAAKASUO *on kognitiotieteen postdoc -tutkija ja Moralities of Intelligent Machines -ryhmän vetäjä.*

JUSSI PALOMÄKI *(PhD) toimii kognitiotieteen tutkijatohtorina Moralities of Intelligent Machines -tutkimusryhmässä Helsingin yliopistossa. Palomäki on robotiikan moraalipsykologian ohella kiinnostunut emootioista, päätöksenteosta ja erityisesti pokeriin liittyvästä tutkimuksesta.*