Kuinka ihmismieli vääristää keskustelua tekoälyn riskeistä ja etiikasta – Kognitiotieteellisiä näkökulmia keskusteluun

MICHAEL LAAKASUO, AKU VISALA JA JUSSI PALOMÄKI

Ihmistämistä kategorioiden välimaastossa

Arkiajattelumme eri alueilla nojaa pääsääntöisesti joukkoon luontaisesti kehittyviä luokitteluja ja kategorisointeja (esim. Atran 2002; Boyer 2001). Kuvaamme ensin tätä kategorisointia kognitiivisena ilmiönä, ja osoitamme sitten, kuinka hankalasti tekoälyn käsite sopii arkiajattelun kategorioihin. Tämä luo tilanteen, jossa tekoälyä voidaan pitää ainakin jossakin määrin intuitionvastaisena käsitteenä. Tekoälyn intuitionvastaisuus selittää monia sen etiikkaan liittyviä virhepäätelmiä ja vääristymiä (Davis & Marcus 2015).

Tarkoitamme termillä *tekoäly* kaikkea sitä teknologiaa, jolla tehdään päätöksiä. Tekoäly ei siis ole sama asia kuin syväoppiva hermoverkko, vaan se kattaa joukon erilaisia ohjelmistoteknisiä ja tilastomenetelmällisiä algoritmisia ratkaisuja usein lokaaleihin tarkasti rajattuihin ongelmiin. Tekoälyä on vaikea määritellä yksiselitteisesti (Warwick 2013). Tämä johtuu pitkälti siitä, että *älykkyyttä* on vaikea määritellä yksiselitteisesti.

Määrittelemme älykkyyden tavoitehakuiseksi ja tarkoituksenmukaiseksi toiminnaksi osittain ennustettavassa ympäristössä (Legg & Hutter 2007). Jos algoritmi tai muu ihmisen tai muun elävän olennon luoma ohjelmistoratkaisu toimii tavoitteellisesti omassa rajatussa ympäristössään, voidaan puhua tekoälystä sanan kapeassa merkityksessä (Warwick 2013).[1]

Arkiteoriat viittaavat ihmisten luontaiseen taipumukseen pyrkiä ennustamaan ja selittämään erilaisia luonnon ilmiöitä ilman erityistä tieteellistä koulutusta (esim. Guglielmo ym. 2009). Arkipsykologia ennustaa toisten ihmisten (tai muiden eläinten) käyttäytymistä, tunteita ja ajatusmaailmaa; arkibiologia puolestaan erilaisten eliöiden kehityshistoriaa ja lisääntymismekanismeja. Taipumuksemme muodostaa intuitiivisia selityksiä ilmiöille on muokkautunut evoluutiohistoriassa. Esi-isämme kehittivät nykytiedon valossa naiiveja selityksiä erilaisille luonnon ilmiöille, kuten maanjäristyksille, vuodenaikojen vaihtelulle, tai taivaankappaleiden liikehdinnälle (esim. Boyer & Barrett 2005). Jotta arkiteoriat voisivat toimia, ihmismielen tulee kyetä luokittelemaan ja kategorisoimaan erilaisia ilmiötä erilaisiin luokkiin. Näihin luokkiin sitten sovelletaan toisistaan poikkeavia, pitkälti automaattisia, päättelyperiaatteita. Luokkia ovat esimerkiksi ”työkalu” (Putt ym. 2017), ”petoeläin” (Boyer & Barrett 2005), ”lemmikki” (Amiot & Bastian 2015) ja ”kasvi” (Atran ym. 2004).

Kategorisoinnin yhtenä tärkeänä rajalinjana on ero toimijoiden, eli agenttien (johon kuuluvat ihmiset ja jotkut eläimet) ja ei-agenttien (kasvit, artefaktit, työkalut jne.) välillä (Barrett 2000). Kyky tehdä tämä erottelu on ollut keskeinen ihmisen selviämisen kannalta. Nykyinen kehityspsykologinen tutkimus on paljastanut, kuinka jo hyvin varhaisessa vaiheessa lapset

1 Viittaamme kirjoituksessamme sekä tieteellisiin lähteisiin että muihin tuoreisiin raportteihin, uutisjuttuihin ja selvityksiin. Tieteellisiin lähteisiin on viitattu perinteisesti, muihin kirjoituksiin alaviitteillä. Tällä tavalla teoreettinen keskustelu on helpompi erottaa arkisista esimerkeistä.

pystyvät erottamaan toimijat esimerkiksi ihmisten valmistamista esineistä eli artefakteista (Johnson 2003). Agentit ovat pääsääntöisesti biologisia olioita, joiden käyttäytymisen selittämiseen arkipsykologiaa voidaan soveltaa. Ei-agenttien selittämiseen puolestaan tarvitaan arkifysiikkaa. Myös erilaisten artefaktien, erityisesti työkalujen, käyttötarkoitusten ymmärtämiseen ihmisellä on luonnollinen kognitiivinen varustus, joka kehittyy varhaisessa ikävaiheessa (Silva & Silva 2006).

Kategorisoinnin kannalta tekoäly muodostaa kuitenkin merkittävän ongelman. Ihmisillä ei ole evoluution muokkaamaa ja luontaista kykyä havaita monimutkaisia informaatiota käsitteleviä järjestelmiä, kuten tekoälyjä tai robotteja, jotka ovat ikään kuin toimijoita mutta samanaikaisesti myös artefakteja. Robotiikan ja kehityspsykologian alan tutkimuksessa on tätä silmällä pitäen esitetty uusi käsite, *uusi ontologinen kategoria,* johon viimeaikainen uusi autonominen teknologia kuuluu (Kahn ym. 2011). Planeettamme historiassa ei ole aikaisemmin ollut tilannetta, jossa elottomasta aineesta rakennettu asia liikkuu, käyttäytyy ja reagoi aivan kuten elävät oliot. On tietenkin totta, että historiamme tuntee suuren joukon erilaisia koneita, joiden tehtävänä on ollut imitoida ihmisten tai eläinten käyttäytymistä. Nämä ”automaatit” ovat kuitenkin olleet hyvin primitiivisiä verrattuna nykyisiin ja lähitulevaisuudessa odotettaviin järjestelmiin (Boden 2006). Sekaannuksien välttämiseksi on tärkeä huomata, että ontologisilla kategorioilla emme viittaa esimerkiksi jossakin kulttuurissa ylläpidettyyn luokitteluun, vaan ihmisen tiedonkäsittelyssä esiintyvään implisiittisiin taipumuksiin, jotka vaikuttavat kaikkien kulttuurien taustalla. Tällä tasolla on selvää, ettei ihmiskognitiolla ole ollut tarpeeksi aikaa kehittää luontaista kykyä, jolla hahmottaa autonomisen teknologian kategoriaa. Nykyään monet tekoälyt tekevät itsenäisesti jopa eettisiä päätöksiä, joilla on suoraan tai välillisesti vaikutus ihmisen hyvinvointiin (Wallach & Allen 2008) – olkoonkin, että päätökset ovat mekanistisia ja tehdään ilman tietoisuutta (ks. Dennett 2003). Kognitiiviset arkiteoriamme ovat kehittyneet ylä-pleistoseeniympäristössä (noin 2 miljoonaa – 200 000 vuotta

sitten) (Tooby & Cosmides 2005). Emme siten ole evoluutiohistoriassamme olleet vuorovaikutuksessa robottien, tietokoneiden, algoritmien tai kyberneettisten järjestelmien kanssa, ja näin meiltä puuttuu arkiteorioihin pohjautuva luontainen kyky kategorisoida niitä ja ennustaa niiden toimintaa.

Robotit ja muut tekoälyt ovat toisin sanoen haaste ihmisen kivikautiselle kognitiolle: riippuen tekoälyjärjestelmän ulkomuodosta luokittelemme ne esimerkiksi eläimiksi, työkaluiksi, leluiksi tai lapsiksi – vaikka ne eivät todellisuudessa ole mitään näistä (Breazeal ym. 2004; Coeckelbergh 2011). Esimerkiksi robottikoiran potkaiseminen saa meidät helposti irvistämään ja tuntemaan jonkinlaista myötätuntoa (Melson ym. 2009; Melson ym. 2009). Samoin sosiaalisen vuorovaikutuksen robotit kuten vanhustenhoidossa käytetty Paro-hylje aktivoivat meissä sosiaaliseen kanssakäymiseen liittyviä positiivisia tunteita (Meacham & Studley 2017).

Kehityspsykologisen tutkimuksen valossa tiedämme, että lapset heijastavat mielellistä kyvykkyyttä esimerkiksi pehmoleluihinsa (Geerdts 2016), ja myös aikuisilla on vastaava taipumus. Maija-Riitta Ollila (2019) kutsuu tätä *ihmistämiseksi* kankean antropomorfismi-sanan sijaan. Keskitymme ihmistämiseen kognitiivisena ja biologisena ilmiönä, joka on kaikkien ihmisten yleinen taipumus. Emme analysoi mahdollisia ihmistämiseen liittyviä eroja kulttuurien välillä. Emme myöskään ota kantaa siihen, että onko ihmistäminen hyvä tai huono asia; pyrimme ainoastaan kuvailemaan ihmistämistä kognitiivisena ilmiönä ja esittelemään mahdollisia eettisiä ongelmia, joita ihmistämistaipumuksesta aiheutuu.

Ihmistämisen ja herkän toimijoidentunnistusjärjestelmän tutkimusta on tehty erityisesti kognitiivisen uskonnontutkimuksen piirissä, jossa näiden järjestelmien on katsottu osittain selittävän yliluonnollisten toimijoiden (kuten jumalien, esiisien ja henkien) suosiota (Barrett 2012). Robotit ja tekoälyjärjestelmät tavallisesti aktivoivat näitä samoja mekanismeja. Vaikeuksia kuitenkin syntyy siitä, että ihmistämisessä roboteille ja muille tekoälyille heijastetaan nimenomaan *inhimillinen* tai *osin*

inhimillinen mieli. Niitä siis pidetään ikään kuin ihmisen kaltaisina kokevina, älykkäinä ja tuntevina olentoina. Näin ihmistäminen johtaa harhaan, sillä kyseisten robottien ja tekoälyjen toimintaperiaatteet poikkeavat merkittävästi ihmisen vastaavista. Robotit ja tekoälyt eivät tunne, järkeile tai tiedosta, vaan niiden kognitiivinen toiminta perustuu algoritmeille ja todennäköisyyslaskennalle. Juuri näitä ihmisen arkikognitio ei kuitenkaan kykene käsittelemään kovinkaan hyvin (Cosmides ym. 2010; Haidt 2001; Moutier ym. 2002; Rode ym. 1999).

Vastaavasti evoluutiopsykologisen tutkimuksen valossa tiedetään, että ihmiset eivät luontaisesti ilman koulutusta ole hyviä loogisessa ja aksiomaattisessa päättelyssä, tai yleisesti hahmottamaan ja arvioimaan matemaattisia todennäköisyyksiä (esim. Rode ym. 1999). Tämä voidaan havaita esimerkiksi kognitiivisissa vääristymissä kuten *uhkapelurin virhepäätelmässä*, jossa toisistaan riippumattomien tapahtumien kuten kolikonheiton lopputulosten todennäköisyydet koetaan toisistaan riippuviksi (esimerkiksi koetaan, että kolmen kruunan jälkeen ”tulisi” tulla klaava) (Rabin & Vayanos 2010), tai esimerkiksi siinä, että ohjelmointitaitojen oppiminen on todella työlästä ja aikaa vievää. Kognitiotieteissä on vuosikymmenten aikana tehdyissä tutkimuksissa havaittu, etteivät ihmiset kykene hahmottamaan ehtolauseita kuten syllogismeja oikein ilman kattavaa koulutusta (Evans 2003; Kellen & Klauer 2019). Tämä on keskeistä, sillä ohjelmointi ja ohjelmoinnin ymmärtäminen perustuu *algoritmeille* eli erilaisten ehtolauseiden monimutkaiselle ketjuttamiselle. Ihmisten luontainen heikkous myös todennäköisyyksien ymmärtämisessä on olennaista, sillä nykyiset koneoppimisalgoritmit rakennetaan todennäköisyyden käsitteen varaan. Esitämme, että koska ihmiset eivät kykene intuitiivisesti, ilman pitkää koulutusta, ymmärtämän todennäköisyyksiä tai algoritmeja, heiltä ei voi myöskään lähtökohtaisesti edellyttää ymmärrystä tekoälyteknologiaan tai siihen liittyviin eettisiin ongelmiin.

On mahdollista, että ajan myötä ihmisille muodostuu uuden ontologisen kategorian (johon robotit ja tekoälyt kuuluvat) mu-

kainen kohtelun luokka, jossa on piirteitä sekä ”eläimen”, ”artefaktin” että ”ihmisen” kategorioista. Tällaisen luokan synty ja kohtelun kehittyminen voi edellyttää kulttuurin muutosta ja vuosia kestävää psykologista harjoittelua. Ehkä ongelmia voitaisiin jatkossa välttää, jos ihmisille tarjottaisiin laajaa ja kattavaa koulutusta, esimerkiksi osana kouluopetusta. Tällöin he oppisivat paremmin säätelemään intuitiivisia reaktioitaan ja ymmärtämään tekoälyjen toimintaperiaatteita.

Saako söpöä robottia pahoinpidellä?

Sosiaalisten robottien kohdalla ihmistämisen ja puutteellisten kategorioiden ongelmat tulevat erityisen hyvin näkyviin. Sosiaalisilla roboteilla on jonkinasteinen ei-tietoinen ymmärrys ihmisten sosiaalisen kanssakäymisen dynamiikasta ja säännöistä, ja ne kykenevät siten toimimaan ja kommunikoimaan ihmisten kanssa erilaisissa tilanteissa. Sosiaaliset robotit tehdään yleensä tarkoituksella ihmisen tunteisiin vetoaviksi tai empatiaa herättäviksi (kuten aiemmin mainittu Paro-hylje), ja ne koetaan usein ”söpöinä ja viattomina”. Roboteista siis tehdään tarkoituksella sellaisia, että niitä ihmistetään mahdollisimman helposti.

Tekoälyn etiikkaa koskevassa keskustelussa on väitelty siitä, missä määrin myötätunnon kaltaiset tunteet ja ihmistävät asenteet ovat asiaankuuluvia (Matthias 2015; Shim & Arkin 2013). Onko kysymyksessä petos tai harhautus, jossa ihminen saadaan tuntemaan myötätuntoa sellaista olioita kohtaan, joka ei kykene siihen vastaamaan eikä edes ymmärtämään sitä? Joihinkin hoivarobotteihin saatetaan jopa muodostaa todellista sosiaalista sitoutumista muistuttava suhde: niille kerrotaan tarinoita, niiden kanssa muistellaan menneitä, niitä silitellään ja niiden kanssa voidaan esimerkiksi itkeä (Sharkey & Sharkey 2012). Tällainen petos olisi merkittävä moraalinen ongelma erityisesti silloin, kun sen kohteena olisivat ihmiset, joiden kognitiiviset kyvyt ovat vaurioituneet, kuten vaikkapa muistisairaan vanhuksen tapauksessa (ibid.; Wachsmuth 2018).

Söpöt ja isosilmäiset hoivarobotit muistuttavat meitä lapsista tai eläinten pennuista (Kringelbach ym. 2016); ne tuntuvat meistä viattomilta ja suojeltavilta ja näitä tunteita on vaikea sivuuttaa. Tällöin robottien suunnittelijat käyttävät hyväkseen yleisinhimillistä taipumusta myötätuntoon. Hoivarobottien käyttö vanhustenhoidossa voi pahimmillaan johtaa myös vanhusten *infantilisaatioon*, eli siihen, että vanhuksiin aletaan suhtautua kuin lapsiin – ikään kuin he kävisivät uudestaan läpi lapsuuttaan leikkiessään leluilla (Sharkey & Sharkey 2012). Tällöin aluksi viattomalta näyttävä ihmisen sosiaalisen kognition manipulaatio voikin osoittautua moraalisesti ongelmalliseksi, jopa vaaralliseksi ihmisille. On siis tärkeä tunnistaa sosiaalisiin robotteihin liittyvä illuusion tai petoksen mahdollinen vaara ja pitää mielessä, etteivät nämä robotit ole tuntevia eivätkä tietoisia olioita.

Vai onko sittenkään kysymys petoksesta? On selvää, ettei esimerkiksi hoivaroboteilla ole sellaisia mielentiloja, joita ihmistävät asenteet niihin liittävät. Vaikka ihmiset suhtautuvatkin positiivisesti sosiaalisiin robotteihin, niitä ei kuitenkaan voida pitää sosiaalisina tai moraalisina toimijoina (Wallach & Allen 2008). Voiko hoivarobotin aikaansaama positiivinen tunne huolenpidosta olla kuitenkin arvokas ja hyödyllinen, vaikka hoivarobotti ei kykenekään minkäänlaiseen tietoiseen myötätuntoon eikä moraaliseen vastuuseen? On näyttöä sen puolesta, että positiivisten tunteiden kokeminen ja negatiivisten tunteiden jakaminen voi edistää hyvinvointia (Meacham & Studley 2017), mutta ei ole selvää, mitkä robottien ”petokseen” pohjautuvan sosiaalisten tunteiden jakamisen pitkäaikaiset vaikutukset ovat. On myös aiheellista pohtia, miten voimme aikaansaada positiivisia tunteita hoivatilanteiden yhteydessä turvautumatta petokseen.

Ihmistäminen ja kategorioiden puute aiheuttaa myös sen, että ihmisten on vaikea käsittää ja suhtautua robottien vaurioitumiseen ja ”pahoinpitelyyn” (Ward ym. 2013). Koska robotteja inhimillistetään, jotkut tilanteet näyttävät robottien kaltoinkohtelulta ja moraalisesti ongelmallisilta, vaikka mitään merkittä-

vää moraalista ongelmaa ei välttämättä olisi (ibid.). Yhdysvaltalainen robotiikan alan yritys Boston Dynamics valmistaa taitavasti ja ihmismäisesti (tai muita eläimiä matkivia) liikkuvia robotteja. Jos näitä robotteja tönitään tai potkitaan, tulkitsemme helposti tilanteen arkipsykologisesti robottien ”kiusaamiseksi”, mikä puolestaan herättää meissä kielteisiä tunteita ja empatiaa robotteja kohtaan (ks. myös Melson ym. 2009; Whitby 2008). Tämän ilmiön voi helposti todeta esimerkiksi katsomalla YouTube:sta Boston Dynamicsin videoita ja lukemalla niiden kommenttikenttiä.

Robottien ”pahoinpitelyä” tarkastelleet tutkijat ovat arvelleet, että ihmiset paheksuvat tekojen sijaan robottia pahoinpitelevää henkilöä, koska tämä näyttäisi olevan luonteeltaan kylmä ja myötätunnoton, vaikkakin esimerkiksi tietokoneen näppäimistöä hakkaavan tai puhelinta suutuksissaan paiskovan henkilön ei tyypillisesti katsota olevan moraalisesti myötätunnoton (Carlson ym. 2019; Riek ym. 2009). Tässä näkyy hyvin se, kuinka robotti luokitellaan huomaamatta jonkinlaiseksi ”kvasi-inhimilliseksi” toimijaksi, vaikka kyseessä on lopulta vain hieman taskulaskinta monimutkaisempi sähköinen laite. Vastaavasti virtuaalihahmojen tappaminen tietokonepeleissä ei tyypillisesti aiheuta samankaltaisia kielteisiä tunteita kuin robottien ”pahoinpitely”. Sekä älypuhelimet että tietokonepelien virtuaalihahmot ovat kuitenkin informaation prosessoinnin näkökulmasta samankaltainen ilmiö kuin ihmismäiset tai söpöt robotit.

Robottien pahoinpitelijöiden moraalinen kauhistelu kuitenkin peittää todellisen moraalisen ongelman. Sen sijaan, että oltaisiin huolissaan robottien puolesta, olisi asiaankuuluvampaa olla huolissaan niiden kanssa vuorovaikutuksessa olevien ihmisten puolesta (Coghlan ym. 2019). On täysin mahdollista, että robottien ihmistämisestä johtuen niiden raaka kohtelu voi luoda ympäristön, jossa niiden kanssa vuorovaikutuksessa olevat ihmiset kehittävät moraalisia paheita. Näin robotteja kohtaan suunnatut asenteet itse asiassa muokkaavatkin niitä ihmi-

siä, jotka näitä asenteita kantavat. Tämä mekanismi toimii riippumatta siitä, millainen moraalinen status roboteilla todellisuudessa on (ibid.).

Jotkut filosofit ovatkin esittäneet, että itsetietoisuutta vailla olevien ja kärsimykseen kykenemättömien eläinten julma kohtelu on haitallista ihmisille itselleen (Amiot & Bastian 2015; Coghlan ym. 2019; Johnson & Verdicchio 2018). Välinpitämätön suhtautuminen ”järjettömien eläinten” kärsimykseen edesauttaa taipumustamme kohdella myös toisia ihmisiä julmasti. Tätä on perusteltu vetoamalla siihen, etteivät ihmisen moraalisen toiminnan kannalta relevantit kognitiiviset kyvyt, kuten moraalinen kuvittelukyky, empatia ja myötätunto, ole automaattisia, vaan edellyttävät harjoitusta ja harjaantumista (MacIntyre 1999). 1900-luvun historia on kuitenkin värittynyt lähinnä julmuudella ja kovasydämisyydellä (Glover 1999; Diamond 1997).

Samainen ilmiö voi muodostua moraalipsykologiseksi ongelmaksi myös tekoälyjen ja robottien kansoittamassa maailmassa (Visala 2020). Kun arkitodellisuus on erilaisten älykkäiden järjestelmien täyttämä, mutta näillä ei kuitenkaan ole moraalisen toimijan statusta, ihmiset saattavat tottua julmuuteen ja välinpitämättömyyteen. Tämä voi hyvinkin muokata ihmisten keskinäisiä suhteita. Jos henkilö tottuu kohtelemaan implisiittisesti ihmistämäänsä robottia kuin artefaktia, saattaa tämä haitata moraalisten kykyjen kehitystä. Koska mielemme kohtelee robotteja ikään kuin ne olisivat eläviä ja tietoisia, voivat niiden kanssa opitut toiminta- ja käyttäytymismallit vaikuttaa kielteisesti myös ihmisten väliseen vuorovaikutukseen. Kun siirrymme kohti tulevaisuutta, joissa seksinuket ja muut robottitoimijat vaikuttavat entistä inhimillisemmiltä, voi niiden ”väkivaltainen” kohtelu vuotaa omaan arkeemme ja siihen, miten kohtelemme toinen toisiamme ihmisinä (idib.). Näistä moraalipsykologisista syistä johtuen ei olisi mielestämme huono ajatus kouluttaa ihmisiä kohtelemaan robotteja tietyssä mielessä moraalisesti, tai jopa velvoittaa jonkinlaista kunnioittavaa kohtelua.

Ihmistäminen ja tekoälyjärjestelmien ideaalinen järkevyys

Ihmistämisen toinen ongelma on eräänlainen *eettinen sokeus* (Palazzo ym. 2012) suhteessa tekoälyjen toimintaan tai niiden koettuun vastuuseen. Eettisellä sokeudella viitataan siihen, etteivät ihmiset ole sosiaalisissa tai ”digi-sosiaalisissa” tilanteissa välttämättä tietoisia niihin liittyvistä eettisistä ongelmista (Bazerman & Tenbrunsel 2011). Esimerkkinä eettisestä sokeudesta toimii paljon julkisuutta saanut tapaus, jossa vuonna 2017 United Airlines -lentoyhtiön lennolta jouduttiin poistamaan väkivalloin matkustaja, joka ei suostunut luopumaan ylivaratusta paikastaan.[2] Koska vapaaehtoisia lennolta lähtijöitä ei löytynyt, tietokonealgoritmi päätti kenet koneesta poistetaan.[3] Itse tilanteessa kukaan lentohenkilökunnan jäsen ei kyseenalaistanut järjestelmän päätöstä eikä suostunut joustamaan lainkaan. Tilanne päättyi väkivaltaisesti vastahakoisen matkustajan poistamiseen koneesta, jonka seurauksena tämä sai aivotärähdyksen ja menetti hampaan. United Airlinesin markkina-arvo laski satoja miljoonia tapauksen jälkeen, ja yhtiö maksoi lopulta matkustajalle merkittävän korvauksen.

Algoritmien tekemiä päätöksiä ja suosituksia on siis voitava kyseenalaistaa, haastaa ja reflektoida harkiten. Tässä tilanteessa algoritmeihin perustuva tunnekylmä järjestelmä nähtiin järkevänä tai ainakin riittävän vahvana auktoriteettina päätöksentekoon. On myös mahdollista, että järjestelmä ihmistettiin, tai pikemminkin ”yli-ihmistettiin”, hyvin spesifillä tavalla: siihen joko tietoisesti tai tiedostamatta heijastettiin jonkinlaista ideaalista rationaalisuutta, jolloin sen auktoriteettiin tehdä sitovia ja järkeviä päätöksiä suostuttiin. Järjestelmän päätös todellisuudessa perustui joihinkin sille opetettuihin tai ohjelmoituihin sääntöihin, jotka eivät ottaneet huomioon sosiaalisen vuorovaikutuksen äärimmäisen merkittäviä seikkoja, kuten mahdollisuutta tilanteen tunnepohjaiseen eskaloitumiseen.

---

2 https://yle.fi/uutiset/3-9561826
3 https://yle.fi/uutiset/3-9869354

Tekoälyjärjestelmien oletettua virheettömyyttä tai ideaalia rationaalisuutta (Palomäki ym. 2012) haittaa erityisen paljon myös se, että näiden järjestelmien toimintaperiaatteet ovat meille pääsääntöisesti näkymättömiä (Castelvecchi 2016). Jo nykyiset tekoälyjärjestelmät, erityisesti koneoppimisalgoritmien päätökset, ovat niin monimutkaisia, ettemme kykene oman kognitiivisen välineistömme avulla ymmärtämään niiden toimintaa (Barratt 2013; Warwick 2013). Eettisessä keskustelussa tämä tunnetaan *mustan laatikon ongelmana* (Introna 2007). Koneoppimisalgoritmi voidaan kouluttaa tiettyä tarkoitusta varten – ja se voi toimia tässä tarkoituksessa pääasiallisesti hyvin – mutta algoritmin päätöksentekoprosessia itsessään on käytännössä mahdotonta seurata. Ongelmaan liittyy laaja eettinen keskustelu, johon emme tässä yhteydessä osallistu. Keskustelu koskee erityisesti ihmisten elämään vaikuttavien päätösten läpinäkyvyyttä (Ollila 2019).

Mustan laatikon ongelmassa on kuitenkin psykologinen puolensa. Aiemmin mainitun ”yli-ihmistämisen” takia meillä on myös taipumus pitää mustassa laatikossa syntyviä päätöksiä ihmisen päätöksiä luotettavampina, koska algoritmit nähdään usein kylmän rationaalisina ja jopa virheettöminä tiedonlähteinä (Lee 2018). Tässä kehityksen vaiheessa on kuitenkin äärimmäisen epätodennäköistä, että ne kykenisivät ottamaan huomioon kaikkea sitä hiljaista (engl. *tacit*) tietoa, joka ihmisten toimintaa jatkuvasti ohjaa (Warwick 2013). Jos joku lentoemännistä tai stuerteista olisi esimerkiksi yrittänyt suostutella tuhansien dollarien palkkiota vastaan (tarjottu palkkio oli n. 800$:n arvoinen) jotain muuta matkustajaa poistumaan koneesta, PR-katastrofilta ja miljoonien menetykseltä oltaisiin mahdollisesti vältytty. Algoritmi ei tähän kuitenkaan kyennyt, ja sen auktoriteettia ei haluttu tai uskallettu kyseenalaistaa.

”Tekevät vain mitä on ohjelmoitu” -vääristymä

Tekoälyjärjestelmien, erityisesti koneoppimisen, ihmiskognition peruskategorioita rikkova luonne aiheuttaa myös toisenlaisia ongelmia. Totesimme edellä, että tekoäly ei selkeästi asetu

niin ”eläimen”, ”artefaktin” kuin ”agentinkaan” kategorioihin. Tällaiset kategorisoinnit kuitenkin ohjaavat intuitiivista ajattelua, joten on syytä tarkastella niitä tässä yhteydessä hieman tarkemmin. Haluamme erityisesti kiinnittää huomiota siihen, mitä tapahtuu jos tekoäly ahdetaan yksinkertaisesti ”artefaktin” ja ”työkalun” kategorioihin. Tällöin on tapana ajatella, että tekoäly (tai laajemmin mikä tahansa ohjelmaa toteuttava tietokone) tekee vain sitä, mitä se on ohjelmoitu tekemään. Tämä oletus, jota tässä kutsumme ”tekee vain mitä on ohjelmoitu” -vääristymäksi, on harhaanjohtava, ja haittaa merkittävästi tekoälyn etiikasta käytyä keskustelua (Lee 2018).

Esimerkiksi erilaiset vahvistusoppimisalgoritmit kykenevät oppimaan hyvin joustavasti ja muokkaamaan omia välitavoitteitaan. Ne eivät siis toteuta jotakin ohjelmoijien ennalta määräämiä päämääriä, vaan kykenevät tunnistamaan niitä itse. DeepMind Technologies -yrityksen tutkijat kehittivät vuonna 2013 vahvistusoppimisalgoritmiin pohjautuvan tekoälyn, joka oppi pelaamaan useita eri Atari-pelejä – joitain paremmin kuin parhaimmat ihmispelaajat (esim. Tegmark 2017). Kyseinen algoritmi sai syötteeksi pelkästään ruudun pikseleitä ja pyrki niiden perusteella lukuisten pelikertojen kautta maksimoimaan pelissä saatuja pisteitä – eli oppimaan pelin pelaamisen. Tekoäly oppi pelistä riippuen tekemään juuri sitä, mikä kyseisessä pelissä auttoi maksimoimaan pisteet. Sillä ei ollut merkitystä, oliko kyseessä esimerkiksi virtuaalihahmojen tappaminen tai lentokoneen lentäminen. Vastaavasti miljardööri Elon Muskin OpenAI-projektin luoma tekoälyohjelma opetteli pelaamaan suosittua *Dota 2* -tietokonepeliä, ja onnistui oppimaan monimutkaisia petoskäyttäytymiseen perustuvia väijytys- ja harhautusstrategioita maksimoidakseen pelipisteensä.[4]

Vahvistusoppimiseen perustuvilla tekoälyillä on kyky tai taipumus löytää odottamattomia ratkaisuja hyvin määriteltyihin ongelmatilanteisiin. Tekoälyt ovat pohjimmiltaan (yksinkertaisia) kognitiivisia toimijoita, jotka oppivat monimutkaisia

4 https://www.wired.com/story/can-bots-outwit-humans-in-one-of-the-biggest-esports-games/

käyttäytymismalleja useissa erilaisissa ympäristöissä. Tekoälyä ei tarvitse erikseen ohjelmoida esimerkiksi tappamaan virtuaalihahmoja, mikäli virtuaalihahmojen tappaminen on jo sidottu osaksi pisteiden maksimointia. Tekoäly tarvitsee vain ohjelmoida maksimoimaan pisteitä, minkä jälkeen se voi kehittyä nopeasti äärimmäisen taitavaksi missä tahansa rajatussa toiminnassa. Tekoälyjen käyttäytyminen voi olla myös ennalta-arvaamatonta, mikäli ne ovat oppineet ihmiselle tuntemattomia käyttäytymismalleja tai strategioita. Näin tapahtui esimerkiksi silloin, kun AlphaGo-algoritmi oppi Go-pelissä ihmiselle tuntemattomia strategioita (Tegmark 2017). Tällaiset tekoälytoimijat voitaisiin teoriassa kouluttaa myös esimerkiksi tappamaan ”ihmisiä” äärimmäisen todenmukaisissa sotasimulaatioissa. Autonomisten pölynimurien ja sotalennokkien testaaminen ja kouluttaminen tapahtuu jo nyt virtuaaliympäristöissä, ja kun testitulokset ovat riittävän hyviä, toimintaa ohjaava koodi voidaan helposti kopioida fyysiseen robottiin.

Lähihistoriamme tarjoaa jo esimerkkejä siitä, kuinka arvaamattomia algoritmit voivat olla. Ajatellaan vaikkapa pörssialgoritmeja, jotka toimivat pääasiallisesti luotettavasti vaihtaessaan arvopapereita ihmisistä koostuvassa ekologisessa ympäristössä (O’Neil 2016). Mikäli kuitenkin samaan pörssiin tuodaan useampi algoritmi, joita ei ole testattu toisiaan ja muita ihmisiä vastaan, algoritmit saattavat muodostaa palautekehiä, joiden seurauksena voi tapahtua odottamattomia, joskin väliaikaisia, pörssiromahduksia. Ainakin kolme tällaista tapausta on jo dokumentoitu.[5] Koska meillä on taipumus ylenkatsoa näitä ongelmia ja olettaa, että tekoäly tekee vain ja ainoastaan sen, mihin se on ohjelmoitu, olisi hyvä kiinnittää erityistä institutionaalista ja laillista huomiota algoritmien testaamiseen ja arviointiin erilaisissa ympäristöissä. Tekoälyteknologian koulutta-

5 https://www.sec.gov/news/studies/2010/marketevents-report.pdf ; https://www.bbc.com/news/business-42959755 ; https://www.theguardian.com/business/2013/apr/23/ap-tweet-hack-wall-street-freefall

minen ja testaaminen tietyssä ympäristössä ei takaa, että tekoäly toimii odotetulla tavalla tai turvallisesti jos ympäristö muuttuu (Hibbard 2012).[6]

Ongelma voidaan esittää myös seuraavasti: Vahvistusoppimiseen perustuvat koneoppimisalgoritmit on ohjelmoitu *oppimaan* (esimerkiksi pisteiden maksimointiin johtavaa toimintaa). Tällöin ne kylläkin tekevät vain ja ainoastaan sitä, mihin ne on ohjelmoitu, eli oppivat; mutta se, *mitä* ne oppivat on täysin ohjelmoidun algoritmin ulkopuolisen ympäristön määrittelemää. Tämä puolestaan johtaa sellaiseen käyttäytymiseen, jota kukaan ei ole tekoälyyn sellaisenaan ohjelmoinut. Kyky kehittää odottamattomia ratkaisuja on syy siihen, miksi oppimisalgoritmit ovat hyödyllisiä, mutta samanaikaisesti myös syy siihen, miksi ne aiheuttavat monimutkaisia eettisiä ongelmia. Juuri oppimisen takia tekoälyn toiminnan ja toimintaympäristön rajaaminen ei ole etukäteen helppoa (Hibbard 2012). Valmiitkin ja huolellisesti testatut algoritmit voivat käyttäytyä koulutusympäristönsä ulkopuolella odottamattomilla tavoilla. Tekoälyltä puuttuvat moraaliset pidäkkeet ja hidasteet, jotka parhaimmassa tapauksessa estävät ihmisiä tekemästä katastrofaalisia moraalisia valintoja. Algoritmin ohjaamalle robotille ei kuitenkaan ole mitään merkitystä, tappaako se virtuaali-ihmisiä tietokonepelissä vai oikeita ihmisiä sodassa.

Tekoäly on moraalisesti relevantti olematta tietoinen

Jos edellä kävi ilmi, millaisia ongelmia seuraa siitä, että tekoäly luokitellaan lähinnä ”artefaktiksi” tai ”työkaluksi”, haluamme seuraavaksi nostaa esiin ongelmia, jotka ovat tulosta siitä, kun tekoälyä pidetään ikään kuin moraalisena toimijana, ”agenttina”. Taipumus pitää tekoälyä moraalisena toimijana on hyvin vahva. Tavallisesti ihmismieli päättelee toimijuuden ytimen,

6 https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-agents-startle-researchers-with-unexpected-strategies-in-hideandseek ; https://www.bbc.com/future/article/20170410-how-to-fool-artificial-intelligence

minuuden ja tietoisuuden olemassaolon älykkään, päämääräsuuntautuneen toiminnan olemassaolosta. Toisin sanoen, jos havaitsemme älykästä toimintaa, oletamme, että toimija on tietoinen – tai ainakin, että toimijalla on jonkinlainen intentionaalinen (esimerkiksi haluava ja tunteva), päämääriin pyrkivä minuus. Tällainen kognitiivinen strategia on hyödyllinen ympäristössä, joka koostuu pääsääntöisesti muista ihmisistä, kuten *Homo sapiensin* ympäristö on jo vuosituhansia koostunut. Kun päämääräsuuntautunutta käytöstä havaitaan, se aktivoi arkipsykologian järjestelmät ja alamme vetää johtopäätöksiä tämän toimijan haluista, ajatuksista, tunteista ja uskomuksista.

Kuitenkin arkipsykologian käyttö muodostuu ongelmaksi silloin, kun sitä sovelletaan tekoälyyn. Tekoälyjärjestelmät voivat nimittäin käyttäytyä siinä mielessä älykkäästi, että ne kykenevät tavoitteelliseen toimintaan jossakin tietyssä ympäristössä. Tästä huolimatta näiltä järjestelmiltä puuttuu tietoisuus: niillä ei ole näkökulmaa omaan itseensä eikä maailmaan; eikä mikään ”tunnu” tai ”vaikuta” niistä miltään (Honkela 2017). Emme tässä yhteydessä halua ottaa vahvaa kantaa laajaan filosofiseen keskusteluun tietoisuudesta ja keinotietoisuuden mahdollisuudesta (esim. Blackmore & Troscianko 2018). Tyydymme huomauttamaan, ettei ainakaan nykyisillä järjestelmillä ole minkäänlaista minuutta tai toimijuutta, puhumattakaan meidän tunnistamastamme tietoisuudesta. Tämä puute ei kuitenkaan sulje pois älykkyyden mahdollisuutta. Järjestelmä voi olla älykäs olematta tietoinen, ja tietoinen olento voi olla tai olla olematta älykäs. Näiltä osin tekoäly ei poikkea eläimistä: vaikkei joku eläin kykenisikään tietoisuuteen, sitä voidaan kuitenkin pitää älykkäänä kognitiotieteen tarkoittamassa mielessä sikäli, kun se kykenee joustavaan ja tavoitteelliseen toimintaan omassa ympäristössään. Kognitiotieteessä on tutkittu jopa kasvien älykkyyttä ja oppimista (Gagliano 2015). Keskeistä on se, että älykkyys ei ole pelkästään ihmisaivojen ominaisuus, ja että älykkään toiminnan ei tarvitse olla tietoista ollakseen älykästä.

Tilannetta monimutkaistaa se, ettei minkäänlaista filosofista taikka tieteellistä konsensusta tietoisuuden luonteesta, synnystä ja toiminnasta ole vielä muodostunut. Väittely on käynyt

kuumana aina 1970-luvulta lähtien (Blackmore & Troscianko 2018). Jotkut filosofit, kuten esimerkiksi Daniel Dennett (1992), ovat esittäneet, että mitään periaatteellista eroa inhimillisen tietoisuuden ja konetietoisuuden välillä ei ole. Tietoisuus on aivoissa tapahtuvaa tiedonkäsittelyä, joten keinotekoinen olio voi aivan hyvin kyetä siihen. Kuvittelemme myös, että minuus ja tietoisuus ovat moraalisen toimijuuden edellytyksiä, mutta mitään kognitiivisesta järjestelmästä ja aivoista erillistä minuutta ei ole tämän näkemyksen mukaan olemassa.

Dennettin edustamalla linjalla on kuitenkin omat kriitikkonsa. Esimerkiksi matemaatikko Roger Penrose (1994) ja neurotutkija Giulio Tononi (2012) ovat esittäneet, että tietoisuudella olisi laskennallisia ominaisuuksia, joita ei voi toteuttaa perinteisellä piipohjaisella mikrosirulla tai niin kutsutulla Turingin kone -arkkitehtuurilla (Penrose 1994). Mikäli he ovat oikeassa, klassiseen laskentaan ja mikrosiruarkkitehtuuriin perustavat teknologiat eivät kykene toteuttamaan tietoisuutta. Tekoäly ei siis nykyisessä, laskennallisessa muodossaan, kykenisi tietoisuuteen, puhumattakaan käsitykseen jonkinlaisesta minuudesta ja moraalisesta toimijuudesta.

Kysymys tekoälyn mahdollisesta moraalisesta toimijuudesta on kuitenkin tietoisuuskysymystä laajempi. Jotkut ovat esittäneet, ettei tietoisuus ole ainoa tärkeä tekijä moraalista toimijuutta arvioitaessa. Moraalinen toimijuus on oikeastaan spektri, jolle erilaiset toimijat voivat asettua. Ehkä riittää se, että tekoäly kykenee käyttäytymiseen, joka joiltakin osin vastaa inhimillistä toimintaa, vaikkei se tietoista toimintaa olisikaan (Wallach & Allen 2008). Tämä funktionaalinen vastaavuus voisi sitten toimia perusteena ainakin jonkinlaiselle moraaliselle kohtelulle. Tällöin tekoäly sijoittuisi moraalisen toimijuuden asteikon alapäähän. Keskustelua on käyty esimerkiksi moraalisen vastuun perusteista (Hakli & Mäkelä 2019). Filosofit ovat kuitenkin olleet hyvin skeptisiä sen suhteen, että tekoälyjärjestelmiä voitaisiin missään mielessä asettaa moraalisen toimijuuden asteikon keskivälille tai yläpäähän (Kauppinen 2020).

Robotit ja tekoälyt voivat siis olla älykkäitä, eli suorittaa tehtäviä ja toimia tavoitteenmukaisesti ennustettavassa ympäristössä. Toisaalta tekoälyt eivät kykene laajamittaiseen moraaliseen toimijuuteen (Wallach & Allen 2008), joten niitä tuskin voidaan lähitulevaisuudessa lukea moraalisen tai laillisen vastuun piiriin. Tästä ei kuitenkaan seuraa se, että tekoälyt olisivat moraalisesti neutraaleja tai moraalisesti merkityksettömiä, koska niiden (älykkäällä) toiminnalla voi olla merkitystä ihmisen hyvinvoinnille. Edes yksinkertainen teknologia, vaikkapa vasara tai pesäpallomaila, ei ole moraalisesti neutraalia, vaan kantaa mukanaan suunnittelijoiden ja tuotantoketjunsa arvoja ja päämääriä (Friedman & Kahn 2002; Roberts 2002). Tämä pitää paikkansa vielä enemmän tekoälyn suhteen, koska se jäsentyy osaksi sosiaalista ja moraalista elämäämme; ja pesäpallomailoista poiketen on luonteensa mukaisesti interaktiivisempaa, autonomisempaa ja sosiaalisempaa.

Älykkäiden laitteiden ja teknologioiden käyttäminen tilanteissa, joissa käytöllä on seurauksia tietoisten olioiden hyvinvoinnille, on arvovalinta ja täten myös moraalisesti merkittävä päätös. Vaikka nykyiset kapeat tekoälyt – eli tekoälyt, jotka on suunniteltu oppimaan tietty toiminta tarkasti rajatussa ympäristössä – eivät ole tietoisia, ne kykenevät itsenäiseen ja jossain määrin luovaan ongelmanratkaisuun. Jos tekoäly tekee itsenäisesti ihmisen hyvinvoinnin kannalta merkittäviä päätöksiä, nämä päätökset heijastavat aina jonkun ihmisen tai ihmisyhteisön näkemystä siitä, mikä on tavoiteltavaa ja moraalista.

Välinesokeus ja vulgaarin seurausetiikan ongelmat

Erilaiset moraalisen kognition aiheuttamat yksinkertaistukset ja vääristymät voivat myös ohjata keskustelua tekoälyn etiikasta umpikujiin ja liiallisiin yksinkertaistuksiin. Kuten filosofit ovat vuosituhantisen väittelynsä aikana huomanneet, eettisistä ongelmista ja etiikan peruskysymyksistä on vaikea muodostaa yhtenäistä, johdonmukaista kantaa. Filosofisen etiikan väittelyt käydäänkin usein yhteensopimattomien normatiivisten eettis-

ten ajattelutapojen, kuten esimerkiksi seurausetiikan, velvollisuusetiikan ja hyve-etiikan, välillä. Näiden väittelyiden ymmärtäminen edellyttää paneutumista ja usein myös filosofista koulutusta. Arkinen moraalinen päättelymme on usein yksinkertaisempaa ja ohjautuu sosiaalisista ja moraalisista tunteistamme käsin (Haidt 2001; 2007). Arkiajattelulle tyypillistä on myös vahva taipumus yksinkertaistaa ja suoraviivaistaa monimutkaisia moraalisia ongelmia. Tämä taipumus aiheuttaa useita ongelmia tekoälyn etiikan yhteydessä.

Eräs normatiivisen etiikan teoria väittää teon moraalisen hyväksyttävyyden riippuvan vain ja ainoastaan teon seurauksista (Darwall 2003). Tämän niin kutsutun s*eurausetiikan* sisällä on lukuisia teorioita ja lähestymistapoja, jotka eroavat sen osalta, miten tekojen seurausten hyötyjä ja haittoja arvioidaan. Seurauseettiset argumentit ovat olleet erityisen vahvasti edustettuina keskustelussa uusista teknologioista ja niiden käyttöönotosta (Gogoll & Müller 2017; Goodall 2014; Schäffner 2018). Esimerkiksi automatisoitua liikennettä puolustetaan väittämällä, että se pelastaa lopulta ihmishenkiä. Siksi olisi hyväksyttävää testata automatisoidun liikenteen teknologiaa julkisessa tilassa ja riskeerata muutama kuolonuhri.

Tällaisessa yksinkertaisessa seurauseettisessä päättelyssä on kuitenkin ilmeisiä ongelmia. Automatisoitu liikenne, vaikkapa autonomiset autot, suunnitellaan ennakolta, ja tämän teknologian suunnittelijat tekevät valintoja siitä, kuka onnettomuustilanteessa on uhrattavissa. Tässä paljastuu seurausetiikkaa kohtaan suunnattu, hyvin tunnettu vastaväite: hyötyjen arviointi on erittäin vaikeaa (kenelle hyötyä tulee, millä hinnalla, mitkä tekijät tulee ottaa huomioon jne.). Lisäksi voidaan huomauttaa, että jos tekojen hyödyn maksimointia sovelletaan suoraviivaisesti tekojen moraaliseen arviointiin, seurausetiikka joutuu jännitteeseen joidenkin moraalisten intuitioidemme kanssa. Moraalinen intuitiomme käsittelee tekojen hyväksyttävyyttä myös esimerkiksi moraalisten normien ja arvojen, ei pelkästään niistä seuraavan hyödyn valossa. Jos esimerkiksi jonkun ihmisen sisäelimet voisivat pelastaa kymmenen muuta ihmistä, voitaisiin tämän ihmisen sisäelinten irrottamista ja uusiokäyttöä pitää

moraalisesti oikeana ratkaisuna, vaikka vastoin tämän henkilön tahtoa.

Vastaavia ongelmia voi syntyä tekoälyn etiikan yhteydessä, jos lyhytaikaiseen hyötyyn tähtäävää seurausetiikkaa sovelletaan sokeasti kiinnittämättä huomiota muihin arvoihin, kuten esimerkiksi ihmisarvoon, velvollisuuksiin ja oikeuksiin. Mikäli vaikkapa autonominen auto suojelee kuljettajaansa onnettomuustilanteessa jalankulkijan kustannuksella, silloin se tosiasia, että jollain oli rahaa ostaa kyseinen auto, muodostuu elämän hintaa määrittäväksi tekijäksi ja tekee auton omistajasta jalankulkijaa tärkeämmän ihmisen. Näin yhtäläinen ihmisarvo voi huomaamatta häipyä vulgaarin seurauseettisen päättelyn edessä; samalla teknologia vauhdittaa eriarvoistumista. Näistä syistä on ensiarvoisen tärkeää, ettei tekoälystä käytävää eettistä keskustelua käydä liian suppeasta eettisestä näkökulmasta.

Seurausetiikan soveltamista tekoälyn kysymyksiin haittaavat myös moraalipsykologien tunnistamat muut kognitiiviset vinoumat. Erästä näistä kutsumme tässä yhteydessä *välinesokeudeksi* (engl. *doctrine of double effect*) (McIntyre 2004; Steinhoff 2018).[7] Ihmiset pitävät tappamista toisten pelastamiseksi moraalisesti hyväksyttävämpänä, mikäli se tehdään nappia painamalla tai giljotiinin narua vetämällä, eli välineellisesti, verrattuna siihen, että kuristaisimme ”omilla käsillämme” toisen hengiltä. Teko on kuitenkin seurauksiensa suhteen täysin sama. Esimerkiksi itseään ajavien autojen moraalia pohdittaessa olemme erittäin alttiita välinesokeudelle: emme havaitse eettisen ongelman vakavuutta, koska auto tai jokin muu väline on osana tapahtumaketjua ja häivyttää henkilökohtaista rooliamme toimijana.

Silloin kun ihminen ajaa autoa ja joutuu ihmishenkiä vaativaan onnettomuuteen, vastuukysymykset ovat melko helposti

7 Tuoreessa analyysissaan, joka julkaistiin *The Journal of Ethics* -lehdessä, Uwe Steinhoff (2018) toteaa doctrine of double effect -ilmiön olevan vääristymä, ja kirjoittaa näin: ”The methodology used by defenders of the DDE or related principles is driven by bias and it is deeply flawed.” Emme kuitenkaan tässä yhteydessä paneudu tähän tulenarkaan keskusteluun sen tarkemmin.

käsiteltävissä. Jos autoa ajanut ihminen aiheutti onnettomuuden, häntä voidaan tietyillä reunaehdoilla helposti rangaista; rankaisu tuntuu muista ihmisistä luonnolliselta ja johdonmukaiselta yleisten moraalikäsitysten valossa. Jos kyseessä on itseohjautuva auto, vastuukysymykset monimutkaistuvat. Onko vastuu tällöin autolla itsellään, auton algoritmien ohjelmoijilla, auton valmistajilla, vai yleisemmin yhteiskunnalla? Ketä tällöin kuuluisi rangaista? Todennäköisesti ainakaan auton valmistaneen yrityksen johtoporrasta ei tulla pitämään välittömästi moraalisesti vastuussa. Voi olla, että auton kehittänyt yritys, jota voitaneen pitää ainakin osittaisessa vastuussa auton toiminnasta, saa korkeintaan sakot huolimattomuudesta. Tämä puolestaan voi johtaa siihen, että ihmishenkiä mitataan enenevässä määrin rahassa, ja että kuolleen omaiset eivät koe saaneensa oikeudenmukaista kohtelua tai kokemusta asian ratkeamisesta (engl. *closure*) (Federico ym. 2016).

Suomen tieverkosto on verorahoilla kustannettua julkista tilaa. Itseohjautuvien autojen yleistyessä niistä voi tulla eräänlaisia yksityisten yritysten testilaboratorioita; vahingon sattuessa seuraukset voidaan kuitata rahalla. Pahimmassa tapauksessa vulgaari seurausetiikka johtaa siihen, että olemme kaikki (ilman suostumustamme) julkisessa tilassa toimiessamme uuden teknologian testaajia, eräänlaisia kolarinukkeja. Kenen tahansa ihmisen hyvinvointi voidaan teknologian edistyksen nimissä vaarantaa, häneltä itseltään mitään kysymättä, jotta teknologista tuotetta voidaan kehittää eteenpäin seurauseettiseen moraaliin vedoten, mutta pohjimmiltaan kuitenkin voiton tavoittelemiseksi. Tällainen laajassa mitassa toteutettu pseudotieteellinen koe tuskin läpäisisi yliopistojen tutkimuseettisten lautakuntien hyväksymisprosessia. Yritysten tuotekehittelyn kohdalla yleinen ilmapiiri tuntuu kuitenkin olevan selvästi sallivampi.

Viimeaikaisissa tutkimuksissa on lisäksi osoitettu, että ihmiset haluaisivat itseohjautuvien autojen toimivan varsin yksinkertaisten seurauseettisten periaatteiden mukaan – eli tavalla, jossa mahdollisimman moni ihmishenki säästyy, vaikka se joskus tarkoittaisi auton kyydissä olevien henkilöiden uhraamista

(Awad ym. 2018; Bonnefon ym. 2016). Lisäksi ihmiset haluaisivat, että muut ihmiset käyttäisivät näillä periaatteilla toimivia autoja, mutta eivät kuitenkaan haluaisi itse olla kyseisten autojen kyydissä (ibid.). Toisin sanoen ihmiset kannattavat seurausetiikkaa siihen saakka, kunnes joutuvat itse uhrautumaan toisten edestä.

Jotta vulgaarin seurausetiikan ja välinesokeuden synnyttämiä riskejä voitaisiin hallita, tarvitaan laajaa konsensusta ja sitoutumista yhtäläiseen ihmisarvoon sekä demokraattisesti hyväksyttyihin sopimuksiin, oikeuksiin ja velvollisuuksiin. Meidän on yhdessä säädeltävä teknologian kehitystä, jotta se palvelisi ihmisen hyvää eikä muodostuisi voitontavoittelussaan hallitsemattomaksi voimaksi, joka pakottaa yksilöt oman edistyksensä välineiksi. Onneksi länsimaisen filosofian, teologian ja poliittisen ajattelun perinteessä on runsaasti resursseja vulgaarin seurausetiikan ongelmien torjuntaan. Oikeusvaltio, jossa yksilöllä on ehdoton arvo, jota ei voi muuttaa rahaksi tai hyödyksi, on ehdottoman tärkeä eettisen teknologian kehittämisen edellytys.

Turvallisuus ei ole ainoa arvo

Keskeinen osa tekoälyn etiikkaa koskee uuden teknologian turvallista kehittämistä ja käyttöä. Tässäkin yhteydessä olisi mielestämme tärkeää, ettei arkiajattelun yksinkertaistuksille annettaisi liikaa valtaa. Turvallisuus on vain yksi arvo muiden joukossa, ja jos se nostetaan muiden yläpuolelle, voivat seuraukset olla hyvinkin ongelmalliset.

Tom Cruisen tähdittämässä elokuvassa *Minority Report* esitetään informaatiota prosessoiva järjestelmä, joka osaa ennakoida mahdollisten murhien ja muiden rikosten tapahtumisen. Kaikki rikokset torjutaan ennakolta ja maailma on näennäisesti täysin turvallinen. Maailma, jossa mitään pahaa ei koskaan tapahdu, vaikuttaa päällisin puolin hyvältä ja toivottavalta. Elokuvassa käsitelty tematiikka koskettaa *väärien positiivisten ongelmaa*: monet ihmiset joutuivat ennaltaehkäisevästi vankilaan,

vaikka he eivät olisikaan oikeasti olleet syyllistymässä rikokseen. Periaatteessa kaikki ihmiset ovat potentiaalisia rikollisia, ja jos kaikki ihmiset laitetaan eristysselleihin, kukaan ei koskaan tee rikoksia. Emme voi kuitenkaan koskaan tietää, olisiko jokin rikos välttämättä tapahtunut, ennen kuin se itse asiassa tapahtuu; voimme vain arvioida sille jonkin todennäköisyyden. Moni päällisin puolin rikoksen valmistelulta näyttävä toiminta voi kuitenkin olla täysin harmitonta (esimerkiksi chilikasvien kasvattaminen lämpölamppujen alla).[8]

Persoonallisuus- ja moraalipsykologian tutkimustiedon valossa ihmiset voidaan sijoittaa jatkumolla perusluonteidensa suhteen, hieman yksinkertaistaen, joko a) avomielisiin ja seikkailullisiin tai b) varovaisiin ja huolellisiin (Aluja ym. 2003; Baer & Oldham 2006; Bouso ym. 2015; Carney ym. 2008; Ebstein ym. 2015; Feist & Brady 2004; Furnham ym. 2009; Hirsh ym. 2010; Ludeke ym. 2013; McCann 2011; Nicholson ym. 2005; Selby ym. 2005; Napier & Luguri 2013; Zeigler-Hill ym. 2015). Avomieliset ja seikkailulliset ihmiset ovat ihmisyhteiskunnissa vähemmistö, mutta tuottavat merkittävän määrän uusista ideoista, keksinnöistä ja oivalluksista. Huolelliset ja varovaiset ihmiset puolestaan vastaavat yhteiskunnan ja organisaatioiden toiminnasta ja ylläpitämisestä. Näihin luonteenpiirteiden ryppäisiin liittyy myös eriävä suhtautuminen arjen moraalisiin kysymyksiin (Clark ym. 2017). Avomieliset ja seikkailulliset ihmiset ovat taipuvaisia arvioimaan moraalisia tekoja ja yhteiskuntia pitkälti siitä näkökulmasta, että ovatko ne i) reiluja tai ii) koituuko tekojen tai poliittisen päätösten seurauksena joillekin vahinkoa. Huolelliset ja varovaiset ihmiset ovat puolestaan taipuvaisempia moraalisissa arvioinneissaan ottamaan huomioon myös muita seikkoja. He arvioivat moraalisia tekoja tai poliittisia päätöksiä suhteessa siihen, kunnioittavatko ne iii) yhteistä tapakulttuuria; iv) yhteiskunnan auktoriteetteja tai v) pyhinä pidettyjä arvoja (olivatpa ne mitä tahansa) (Graham ym. 2011; Haidt ym. 2009). Varovaiset ja huolelliset ihmiset ovat tavallisesti

8 https://www.oikeusasiamies.fi/r/fi/ratkaisut/-/eoar/4073/2009

myös herkempiä kokemaan inhon ja pelon tunteita kuin avoimet ja seikkailulliset ihmiset (van Leeuwen ym. 2017).

Näistä luonne-eroista johtuen monia tekoälysovelluksia ja tulevaisuuden teknologioita saatetaan markkinoida ihmisille pelotteluun ja turvallisuuteen nojaten (Bird & Tapp 2011; Brennan & Binney 2010; Mohr ym. 2010). Tällaisia teknologioita ovat esimerkiksi kasvojentunnistusalgoritmeihin perustuva ihmisten automaattinen tunnistaminen ja profilointi, ja Yhdysvalloissa jo nyt käytössä olevat ohjelmat, jotka pyrkivät ennakoimaan niitä kaupunginosia, joissa seuraava mahdollinen rikos saattaa tapahtua.[9] Ohjaamalla poliisit kyseiseen kaupunginosaan, tästä algoritmista voi tulla *Minority Report* -elokuvan esittämällä tavalla itsensä toteuttava ennuste: poliisit pyrkivät löytämään kenet tahansa ihmisen, jolla saattaisi olla kannabista taskussaan, koska kyseisen alueen vastaavien rikosten todennäköisyys on suurempi. Järjestelmä ei kuitenkaan sovellu yhtä hyvin talousrikosten ehkäisemiseen, vaikka talousrikollisten aiheuttamat vahingot ovat yhteiskunnalle kokonaisuuden kannalta vakavammat. Talousrikoksia ei suoraan tai luontaisesti nähdä turvallisuusuhkina, vaikka ne saattavat pitkällä aikavälillä – yhteiskunnan heikentyneen rahoituspohjan myötä – ilmetä kasvaneina itsemurhalukuina, alkoholismina ja perheväkivaltana (Eubanks 2017).

Ilman koulutusta ihmisten on vaikea ymmärtää tekoälyn etiikkaa tai siihen liittyviä sosiologisia ongelmia; ja asiaan vihkiytymätön väestö voi olla helposti suostuteltavissa turvallisuuteen vedoten ottamaan käyttöön sellaista teknologiaa, joka saattaa rapauttaa perustuslaillisen demokratian toimintaedellytyksiä.

Meidän ja muiden rikokset

Ihmisen moraalisen ja sosiaalisen mielen evoluutio on tapahtunut suhteellisen pienten ja keskenään ainakin jossakin määrin kilpailutilanteessa olevien ryhmien kontekstissa (Boyd &

9 https://www.technologyreview.com/s/612957/predictive-policing-algorithms-ai-crime-dirty-data/

Richerson 2005; Sober & Wilson 1998). Ihmismielen erityisyys on juuri sen läpeensä sosiaalisessa luonteessa (Gronow & Kaidesoja 2017). Eräs sosiaalisen mielemme taipumus onkin tehdä ero sisäryhmän ja ulkoryhmän välillä (Fiske & Taylor 2008). Tästä jaottelusta seuraa se, että sisäryhmään kuuluvat yksilöt on helpompi nähdä arvokkaampina kuin ulkoryhmään kuuluvat. Voisimme kutsua tätä *me ja muut -vääristymäksi*. Tämä vääristymä vaikuttaa myös tekoälyn etiikkaa koskevan keskustelun alueella.

Lähihistoriasta löytyy ainakin yksi esimerkki, jossa tekoälyalgoritmia hyödynnettiin EU:n sisällä veronkierron ja muiden talousrikosten torjunnassa (Varoufakis 2016; 2017). Kun algoritmi oli saatu valmiiksi, se ajettiin tietyn valtion verotietokannan läpi. Tietokannasta oli tarkoitus poimia tuhansia tarkkailuun ja tutkinnan alle laitettavia nimiä, mutta Euro-ryhmä ja troikka[10] estivät tämän toimeenpanon (ibid.). Vastaavanlaista EU:n huipputason puuttumista rikosten torjuntaan tuskin olisi tapahtunut paikallisella tasolla, jos vastaavia suuria aineistoja olisi käytetty esimerkiksi paikallisten ihmisten huumeiden käytön torjumisessa. Uusi teknologia kuitenkin mahdollistaa uudenlaisen tiedon urkinnan ja rikosten torjunnan; olisi teoriassa mahdollista luoda tietokanta, joka yhdistää ihmisten musiikkitottumukset heidän matkustustietoihinsa ja asuinalueensa jätevesien kemialliseen analyysiin (jonka perusteella voidaan analysoida huumausaineiden käyttöä) tai muihin vastaaviin tietoihin.

Suomessa on satoja tuhansia laittomia päihdeaineita käyttäviä ihmisiä; ja huumausaineiden käyttö on viime vuosina lisääntynyt merkittävästi.[11] Yksi stereotypia monien päihdeaineiden kohdalla lienee, että niitä käyttävät vain syrjäytymisvaa-

10 https://en.wikipedia.org/wiki/European_troika
11 https://thl.fi/tilastot-ja-data/tilastot-aiheittain/paihteet/huumeet/suomalaisten-huumeiden-kaytto-ja-huumeasenteet; https://thl.fi/-/jatevesitutkimus-amfetamiinia-kaytetaan-ennatyksellisen-paljon-myos-kokaiinin-kaytto-lisaantynyt-edelleen

rassa olevat köyhät. Kansainvälisistä vertailevista tutkimuksista kuitenkin tiedetään, että suurin osa päihteitä käyttävistä ihmisistä ei koskaan jää kiinni eikä heidän päihteiden käytöstään ole havaittavissa olevaa haittaa muulle yhteiskunnalle (Müller & Schumann 2011). Laittomia päihdeaineita käyttävät myös yhteiskunnan hyväosaiset jäsenet kuten lääkärit, asianajajat, kliiniset psykologit, professorit, IT-yrittäjät, fyysikot, peruskoulun opettajat ja sosiaalityöntekijät (ibid.). Mikäli valtion hallinnolla olisi tahtoa, voitaisiin tuhansien veronmaksajien elämää merkittävästi hankaloittaa tämänkaltaisen arkaluontoisen informaation selvittämisellä.

Näissä yrityksissä näkyy varsin hyvin me ja muut -vääristymä: oma sisäryhmä ja sen tottumukset koetaan helposti arvokkaampina ja tärkeämpinä kuin ulkoryhmän tottumukset (esim. Voci 2006). Tämä korostuu etenkin siinä, kuinka tekoälysovelluksia rikosten torjunnassa ollaan herkempiä käyttämään sellaisia ryhmiä kohtaan, joihin päättäjät eivät itse kuulu, kuten köyhiin tai maahanmuuttajiin[12,13]. Ottaen huomioon yllä kuvatut erot yksilöiden moraaliarvioiden lähtökohdissa (suurin osa ihmisistä miettii arjessaan tekojen oikeutusta suhteessa siihen, kunnioittaako teko oman kulttuurin arvoja), saattavat ihmiset tukea myös profilointialgoritmien käyttöönottoa rikosten torjunnassa. Ihmisten on paljon vaikeampi kuvitella, että he olisivat itse osa jotain ulkoryhmää ja joutuisivat valvontateknologian tarkkailun kohteeksi. Uutta ”turvallisuutta lisäävää” teknologiaa ja sen käyttöönottoa voidaan pitää hyvänä ideana – kunnes joudumme itse sen uhriksi.

Toisaalta, jos etniseen tai muuhun profilointiin kykenevä valvontakoneisto saadaan luotua, pystytään se helposti kalibroimaan uudelleen uusien kohteiden tarkkailuun. Onko hyvä,

12 https://www.theguardian.com/technology/2019/oct/16/digital-welfare-state-big-tech-allowed-to-target-and-surveil-the-poor-un-warns ; https://www.wired.com/story/opinion-ai-for-good-is-often-bad/

13 https://www.cigionline.org/articles/using-ai-immigration-decisions-could-jeopardize-human-rights

että ihmisten toimintaa kontrolloidaan ja valvotaan näin tarkasti? Yhteiskunnan liikkuminen eteenpäin eetoksensa ja moraaliymmärryksemme suhteen on usein sidoksissa ”harmaalla alueella” tapahtuviin innovaatioihin (Brownlee 2017). Vastikään esimerkiksi HUSin eettinen toimikunta on antanut luvat Suomessa psilosybiinitutkimuksille, joissa on tarkoitus hoitaa erittäin vakavasti masentuneita ihmisiä.[14] Tutkimusidea ja siihen liittyvä taustatyö on osiltaan ollut yhteiskunnassa marginaaliin jäävien aktivistien mahdollistamaa (henkilökohtainen tiedonanto) ja samainen aktivistiverkosto on myös antanut lausuntoja korkeimmalle oikeudelle, joka muutti niin kutsuttujen taikasienten vaarallisuusasteen miedompaan kategoriaan.

Egosentrinen teleologiavääristymä

Egosentrisyys viittaa ihmisen tapaan hahmottaa maailmaansa omasta näkökulmastaan. Jokainen ihminen kokee olevansa erikoinen ja erityinen, ja ymmärtää olevansa ainutlaatuinen ja melko monimutkainen toimija (Preston 2018). Ihmisellä on taipumuksena nähdä myös ulkopuolinen maailmansa siten, että se on ikään kuin olemassa häntä varten. Egosentrinen näkökulma heijastuu osittain myös *teleologiseen* tapaan havainnoida maailmaa: tuolit on suunniteltu istumista varten ja keihäät metsästystä varten (ja molemmat ovat siksi ”hyviä”). Maailmassa näyttää olevan suunnitelmallisuutta (ibid.).

Ihmiskunnan historiassa egosentriseen teleologiavääristymään on liittynyt myös se, että toisten eläinten on koettu olevan olemassa ihmisiä varten; esimerkiksi joko ruoaksi tai aputyövoimaksi (ibid.). Egosentrisyysharha voi olla myös osa kokonaista maailmankuvaa, jossa koko todellisuuden, erityisesti muiden eläinten ja kasvien, nähdään olevan olemassa ihmistä varten ja hänen palveluksessaan. Egosentrisyysharha ilmenee myös siten, että ihminen kokee itsensä olevan päämääräsuuntautunut, monimutkainen ja hyvä, joten kaikki ne ryhmät, joihin ihminen itse kuuluu, ovat myös hyviä; eihän hyvä ihminen

14 https://www.hs.fi/tiede/art-2000005382691.html

voi kuulua pahaan ryhmään, ”enkä varsinkaan minä”. Tämän seurauksena ihmiskunnan historiassa on tapahtunut useita kertoja *dehumanisaatioksi* nimitetty ilmiö (Haslam 2006; Haslam & Loughnan 2014; Waytz ym. 2010); orjat eivät ole ihmisiä ja Tutsit ovat torakoita (sic; Rothbart & Barlett 2008).

Ihmisillä on myös tapana nähdä kaikki monimutkaiset ja päällisin puolin päämääräsuuntautuneet teknologiat eettisesti vähemmän ongelmallisina kuin mitä ne saattavat olla. Älypuhelimet, autot, televisiot, ydinvoimalat, profilointialgoritmit, ja esimerkiksi sairaanhoitorobotit saatetaan kokea näennäisesti moraalisesti neutraaleina tai jopa moraalisesti hyvinä asioina. Autoista ja älypuhelimista ei esimerkiksi puhuta sinänsä moraalisina teknologisina tuotoksina, vaikka niiden seurauksista ja käyttötavoista voidaankin puhua moraalisista näkökulmista. Tosiasiassa teknologisia välineitä on mahdotonta rakentaa moraalisessa tyhjiössä. Kuten monet geologit ovat todenneet, olemme siirtyneet antroposeeniin (Oldfield ym. 2014), eli uuteen geologiseen aikakauteen, jossa maapalloa muokataan aktiivisesti ihmisen tarpeisiin.

Egosentrinen teleologiavääristymä on yksi monista tekijöistä, jotka saavat meidät moraalisesti sokeiksi uuden teknologian haasteiden suhteen. Tunnistamme robotit jokseenkin autonomisina päämääräsuuntautuneina toimijoina, mutta samalla ne näyttävät toimivan kuin taikavoimilla: ne ovat niin monimutkaisia, että emme ymmärrä, miksi ne toimivat kuten toimivat. Saatamme siis nähdä päämääräsuuntautuneen monimutkaisen teknologian moraalisesti neutraalina tai hyvänä, vaikka näin ei välttämättä ole. Tämä taas juontuu siitä, että olemme itse päämääräsuuntautuneita, monimutkaisia ja pidämme itseämme ”hyvinä”. Koska uusi tekoälyteknologia on ihmisten luomaa, emme ole luontaisesti taipuvaisia kyseenalaistamaan sen moraalisia seurauksia. Monimutkainen ja uusi teknologia tarvitsee tuekseen hidasta ja selkeää filosofista ja eettistä pohdintaa.

Samalla ei ole pelkästään niin, että koska ajattelemme itsemme hyviksi tai rationaalisiksi, olisivat teknologiset järjestelmät tätä. Näitä kuvitellaan myös siksi, että ne teoreettiset mallit,

joihin järjestelmät nojaavat, nähdään ”rationaalisina” tai todellisuutta hyvin kuvaavina. Esimerkiksi eräät peliteoreettiset mallit ovat toimineet pohjana yhteiskunnallisten instituutioiden toimintojen automatisaation suunnittelussa. On kuitenkin hyvä huomata, että tällaiset mallit ovat rajoittuneita kuvaamaan esimerkiksi ihmisen sosiaalisuutta.[15]

Lopuksi

Halusimme tai emme, olemme ihmiskuntana jo selvästikin siirtyneet uudenlaiseen aikaan, jossa meitä kohtaavat uudenalaiset moraaliset haasteet, jotka ovat sekä syvä- että pintarakenteeltaan uudenlaisia (esim. Harari 2016). Olemme ensimmäistä kertaa ihmiskunnan historiassa tilanteessa, jossa ennalta suunniteltu eloton ja tiedostamaton materiaali tekee päätöksiä ihmisten hyvinvoinnista. Oman kognitiomme rajoitteet ja vääristymät saattavat kuitenkin viedä teknologian kehitystä suuntaan, jossa sen käytännön seuraukset eivät ole demokraattisen ja vapaan yhteiskunnan näkökulmasta haluttavia.

Näiden muutosten tapahtuessa teknologia alkaa toimia oman moraalisen kognitiomme ja poliittisten järjestelmiemme peilinä. Voimme joko tietoisesti varautua suuntaamaan teknologian kehitystä kohti niitä suuntia, jotka vahvistavat demokraattisten järjestelmien, tasa-arvon ja oikeudenmukaisuuden periaatteita; tai voimme antautua teknologian kehityksen prosessille ja antaa sen mukautua omaan kivikautiseen kognitioomme. Oma moraalinen kognitiomme on sekä emootioiden että primitiivisten laumakäyttäytymisvaistojen muovaamaa. Mikäli haluamme välttää sen, että eläinluontomme tarpeet päätyvät ohjaamaan yhteiskuntiemme kehitystä, on teknologian kehityksen motivaatiotekijöitä ja mahdollisia kivikautisen luontomme vääristäviä tekijöitä otettava jo tekoälyjen suunnitteluvaiheessa paremmin huomioon.

---

15 Kiitämme arvioijaa hyvästä kommentista.

Olemme tässä kirjoituksessa käyneet laskutavasta riippuen läpi noin 13 erilaista ihmiskognition piirrettä, jotka osaltaan selittävät sitä miksi tekoälyteknologioista ja niiden riskitekijöistä keskusteleminen on vaikeaa. Yllä esitelty ongelmavyyhti kumpuaa siitä tosiseikasta, että ihminen on kehittynyt evoluutiohistoriassaan ympäristössä, jossa ei ole ollut tekoälytoimijoita, tai todennäköisyyden ja ehtolauseiden käsitteitä. Ihminen on intuitiivista ja automaattista kognitiotaan käyttävä sosiaalinen eläin, jolle on luontaista inhimillistää ja projisoida tietoista toimijuutta ympäristöönsä, omista egosentrisistä lähtökohdistaan käsin. Tekoälyteknologia ei kuitenkaan ole sellaista, että se olisi tämän automaattisen kognition ymmärrettävissä.

Toivomme, että kirjoituksemme antaa uusia pohdinnan aiheita aihealueen parissa työskenteleville ihmisille ja vie yhteiskunnassamme käytävää tekoälyihin liittyvää riskikeskustelua sekä pidemmälle että uusiin suuntiin. Vaihtoehtoina näyttää olevan valistunut ja varovainen keskustelu, jossa riskeihin varaudutaan tietoisesti ja viedään ihmiskuntaa kohti demokraattisempaa ja tasa-arvoisempaa yhteiskuntaa (kuten *Star Trek* -tieteissarjassa); tai se, ettei riskikeskustelua käydä sivistyneesti ja päädytään tilanteeseen, joka muistuttaa enemmän niitä lukuisia dystopioita, joita esimerkiksi *Blade Runnerin* ja *The Matrixin* kaltaisissa tieteisfiktioissa kuvataan.

Helsingin yliopisto

Kirjallisuus

Aluja, Anton, García, Oscar ja García, Luis F. (2003) ”Relationships Among Extraversion, Openness to Experience, and Sensation Seeking”, *Personality and Individual Differences* 35 (3), 671–680.

Amiot, Catherine E. ja Bastian, Brock (2015) “Toward a Psychology of Human–Animal Relations”, *Psychological Bulletin* 141 (1), 6–47.

Atran, Scott, Medin, Douglas ja Ross, Norbert (2004) “Evolution and Devolution of Knowledge: A Tale of Two Biologies”, *Journal of the Royal Anthropological Institute* 10 (2), 395–420.

Awad, Edmong, Dsouza, Sohan, Kim, Richard, Schulz, Jonathan, Henrich, Joseph, Shariff, Azim, Bonnefon, Jean-François ja Rahwan, Iyad (2018) "The Moral Machine Experiment", *Nature* 563 (7729), 59–64.

Baer, Markus ja Oldham, Greg R. (2006) "The Curvilinear Relation Between Experienced Creative Time Pressure and Creativity: Moderating Effects of Openness to Experience and Support for Creativity", *Journal of Applied Psychology* 91 (4), 963–970.

Barratt, James (2013) *Our Final Invention*. Macmillan.

Barrett, Justin L. (2000) "Exploring the Natural Foundations of Religion", *Trends in Cognitive Sciences* 4 (1), 29–34.

Barrett, Justin L. (2012) *Born Believers: The Science of Children's Religious Belief*. New York: The Free Press.

Bird, Sara ja Tapp, Aalan (2011) *Fear and Fire: Ethical Social Marketing Strategies for Home Fire Safety for Older People*. "https://uwe-repository.worktribe.com/output/963462"

Blackmore, Susan ja Troscianko, Emily (2018) *Consciousness: An Introduction, Third Edition*. Lontoo: Routledge.

Bonnefon, Jean-François, Shariff, Azim ja Rahwan, Iyad (2016) "The Social Dilemma of Autonomous Vehicles", *Science* 352 (6293), 1573–1576

Boden, Margaret (2008) *Mind as Machine: A History of Cognitive Science.* Oxford University Press.

Bouso, José Carlos, Palhano-Fontes, Fernanda, Rodríguez-Fornells, Antoni, Ribeiro, Sidarta, Sanches, Rafael, Crippa, Jose A., Hallak, Jaime, Barros de Araujo, Draulio ja Riba, Jordi (2015) "Long-Term Use of Psychedelic Drugs is Associated with Differences in Brain Structure and Personality in Humans", *European Neuropsychopharmacology* 25 (4), 483–492.

Boyer, Pascal (2001) *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.

Breazeal, Cynthia, Gray, Joanna, Hoffman, Geoff, ja Berlin, Matt (2004) "Social Robots: Beyond Tools to Partners", *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 551–556.

Boyer, Pascal ja Barrett, Clark (2005) "Evolved intuitive ontology: Integrating neural, behavioral and developmental aspects of domain-specificity" teoksessa D.M. Buss. (toim.), *Handbook of evolutionary psychology*. Hoboken, NJ: Wiley & Sons, 161–179

Brennan, Linda, ja Binney, Wayne (2010) ”Fear, Guilt, and Shame Appeals in Social Marketing”, *Journal of Business Research* 63 (2), 140–146.

Brownlee, Kimberleyh (2017) “Civil Disobedience”, *The Stanford Encyclopedia of Philosophy,* toim. E. N. Zalta. <https://plato.stanford.edu/archives/fall2017/entries/civil-disobedience/>.

Carlson, Zachary, Lemmon, Louise, Higgins, MacCallister, Frank, David, Salek Shahrezaie, Roya ja Feil-Seifer, David (2019) “Perceived Mistreatment and Emotional Capability Following Aggressive Treatment of Robots and Computers”, *International Journal of Social Robotics* 11, 727–739.

Carney, Dana R, Jost, John T., Gosling, Samuel D. ja Potter, Jeff (2008) “The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind”, *Political Psychology* 29 (6), 807–840.

Castelvecchi, Davide (2016) “Can We Open the Black Box of AI?” *Nature News* 538 (7623), 20.

Clark, Charles Brendan, Swails, Jeffrey, Pontinen, Heidi M., Bowerman, Shannon, Kriz, Kenneth A., ja Hendricks, Peter S. (2017) “A Behavioral Economic Assessment of Individualizing Versus Binding Moral Foundations”, *Personality and Individual Differences* 112, 49–54.

Coeckelbergh, Mark (2011) “Humans, Animals, and Robots: A Phenomenological Approach to Human-Robot Relations”, *International Journal of Social Robotics* 3 (2), 197–204.

Coghlan, Simon, Vetere, Frank, Waycott, Jenny ja Barbosa Neves, Barbara (2019) “Could Social Robots Make Us Kinder or Crueller to Humans and Animals?”, *International Journal of Social Robotics* 11 (5), 741–751.

Cosmides, Leda, Barrett, Clark ja Tooby, John (2010) “Adaptive Specializations, Social Exchange, and the Evolution of Human Intelligence”, *Proceedings of the National Academy of Sciences*, 107 (Supplement_2), 9007–9014.

Darwall, Stephen (2003) *Consequentialism*. Oxford: Blackwell.

Davis, Ernest ja Marcus, Gary (2015) “Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence”, *Communications of the ACM* 58 (9), 92–103.

Dennett, Daniel (1992) *Consciousness Explained*. Penguin UK.

Dennett, Daniel (2003) *Freedom Evolves.* Penguin UK.

Diamond, Jared (1997) *Guns, Germs, and Steel*. New York: WW Norton.

Ebstein, Richard P., Monakhov, Mikhail V., Lu, Yunfeng, Jiang, Yushi, Lai, Poh San ja Chew, Soo Hong (2015) "Association Between the Dopamine D4 Receptor Gene Exon III Variable Number of Tandem Repeats and Political Attitudes in Female Han Chinese", *Proceedings of the Royal Society B: Biological Sciences* 282 (1813), 2015.1360.

Evans, Jonathan S. B. (2003) "In Two Minds: Dual-Process Accounts of Reasoning", *Trends in Cognitive Sciences* 7 (10), 454–459.

Eubanks, Virginia. (2017) *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Federico, Christopher M., Ekstrom, Pierce, Tagar, Michal Reifen ja Williams, Allison L. (2016) "Epistemic Motivation and the Structure of Moral Intuition: Dispositional Need for Closure as a Predictor of Individualizing and Binding Morality", *European Journal of Personality* 30 (3), 227–239.

Feist, Gregory J. ja Brady, Tara R. (2004) "Openness to Experience, Non-Conformity, and the Preference for Abstract Art", *Empirical Studies of the Arts* 22 (1), 77–89.

Fiske, Susan T. ja Taylor, Shelley E. (2013) *Social Cognition: From Brains to Culture*. Sage.

Furnham, Adrian, Crump, John, Batey, Mark, & Chamorro-Premuzic, Tomas (2009) "Personality and Ability Predictors of the 'Consequences' Test of Divergent Thinking in a Large Non-Student Sample", *Personality and Individual Differences* 46 (4), 536–540.

Friedman, Batya, Kahn Jr., Peter.H. (2002) "Human values, ethics, and design" teoksessa J. Jacko (toim.), *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications.* CRC press, 1209–1233.

Gagliano, Monica (2015) "In a Green Frame of Mind: Perspectives on the Behavioural Ecology and Cognitive Nature of Plants", *AoB PLANTS*, 7, doi: 10.1093/aobpla/plu075

Geerdts, Megan S. (2016) "(Un)Real Animals: Anthropomorphism and Early Learning About Animals", *Child Development Perspectives* 10 (1), 10–14.

Glover, Jonathan (1999) *Humanity*. Yale University Press.

Gogoll, Jan ja Müller, Julian F. (2017) "Autonomous Cars: In Favor of a Mandatory Ethics Setting", *Science and Engineering Ethics* 23 (3), 681–700.

Goodall, Noah J. (2014) "Machine ethics and automated vehicles" teoksessa G. Meyer ja S. Beiker (toim.), *Road vehicle automation*. Springer International Publishing, 93–102

Graham, Jesse, Nosek, Brian A., Haidt, Jonathan, Iyer, Ravi, Koleva, Spassena ja Ditto, Peter H. (2011) ”Mapping the Moral Domain”, *Journal of Personality and Social Psychology* 101 (2), 366–385.

Gronow, Antti ja Kaidesoja, Tuukka (2017) *Ihmismielen Sosiaalisuus*. Helsinki: Gaudeamus.

Guglielmo, Steve, Monroe, Andrew E. ja Malle, Bertram F. (2009) “At the Heart of Morality Lies Folk Psychology”, *Inquiry* 52 (5), 449–466.

Haidt, Jonathan (2001) “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment”, *Psychological Review* 108 (4), 814–834.

Haidt, Jonathan (2007) “The New Synthesis in Moral Psychology”, *Science* 316 (5827), 998–1002.

Haidt, Jonathan, Graham, Jesse ja Joseph, Craig (2009) ”Above and Below Left–Right: Ideological Narratives and Moral Foundations”, *Psychological Inquiry* 20 (2–3), 110–119.

Hakli, Raul ja Mäkelä, Pekka (2019) “Moral Responsibility of Robots and Hybrid Agents”, *The Monist* 102 (2), 259–275.

Haslam, Nick (2006) “Dehumanization: An Integrative Review”, *Personality and Social Psychology Review* 10 (3), 252–264.

Haslam, Nick ja Loughnan, Steve (2014) “Dehumanization and Infrahumanization”, *Annual Review of Psychology* 65 (1), 399–423.

Honkela, Timo (2017) *Rauhankone: Tekoälytutkijan Testamentti*. Helsinki: Gaudeamus.

Hibbard, Bill (2012) ”Avoiding unintended AI behaviors” teoksessa J. Bach, B. Goertzel ja M. Iklé (toim.), *Artificial general intelligence*. New York: Springer, 107–116.

Hirsh, Jacob B., DeYoung, Colin G., Xiaowen Xu ja Peterson, Jordan B. (2010) “Compassionate Liberals and Polite Conservatives: Associations of Agreeableness with Political Ideology and Moral Values”, *Personality and Social Psychology Bulletin* 36 (5), 655–664.

Introna, Lucas (2007) “Maintaining the Reversibility of Foldings: Making the Ethics (Politics) of Information Technology Visible”, *Ethics and Information Technology,* 9 (1), 11–25.

Johnson, Deborah G. ja Verdicchio, Mario (2018) “Why Robots Should Not Be Treated Like Animals”, *Ethics and Information Technology* 20 (4), 291–301.

Johnson, Susan C. (2003) “Detecting Agents”, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358 (1431), 549–559.

Kahn, Peter H., Reichert, Aimee L., Gary, Heather E., Kanda, Takayuki, Ishiguro, Hiroshi, Shen, Solace, Ruckert, Jolina H. ja Gill, Brian (2011) "The New Ontological Category Hypothesis in Human-Robot Interaction", *Proceedings of the 6th International Conference on Human-Robot Interaction*, 159–160.

Kauppinen, Antti (Ilmestyy 2021) "Osaammeko rakentaa moraalisia toimijoita?" teoksessa P. Raatikainen (toim.), *Tekoäly, ihminen ja yhteiskunta*. Helsinki: Gaudeamus.

Kellen, David ja Klauer, Karl Christoph (2019) "Theories of the Wason Selection Task: A Critical Assessment of Boundaries and Benchmarks", *Computational Brain & Behavior*, 1–13.

Kringelbach, Morten L., Stark, Eloise A., Alexander Catherine, Bornstein, Marc H. ja Stein, Alan (2016) "On Cuteness: Unlocking the Parental Brain and Beyond", *Trends in Cognitive Sciences* 20 (7), 545–558.

Lee, Min Kyung (2018) "Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management", *Big Data & Society* 5 (1), doi: 10.1177/2053951718756684

Ludeke, Steven, Johnson, Wendy ja Bouchard, Thomas J. (2013) "Obedience to Traditional Authority: A Heritable Factor Underlying Authoritarianism, Conservatism and Religiousness", *Personality and Individual Differences* 55 (4), 375–380.

MacIntyre, Alasdair (1999) *Dependent Rational Animals: Why Human Beings Need the Virtues.* Chicago: Open Court.

Matthias, Andreas (2015) "Robot Lies in Health Care: When Is Deception Morally Permissible?", *Kennedy Institute of Ethics Journal* 25 (2), 169–162.

McCann, Stewart (2011) "Conservatism, Openness, and Creativity: Patents Granted to Residents of American States", *Creativity Research Journal* 23 (4), 339–345.

McIntyre, Alison (2004) "Doctrine of Double Effect", *The Stanford Encyclopedia of Philosophy,* toim. E. N. Zalta. <https://stanford.library.sydney.edu.au/entries/double-effect/>.

Meacham, Darian ja Studley, Matthew (2017) "Could a robot care? It's all in the movement" teoksessa P. Lin, K. Abney ja R. Jenkins (toim.), *Robot ethics 2.0: From autonomous cars to artificial intelligence.* New York: Oxford University Press, 98–111.

Melson, Gail F., Kahn Jr., Peter H., Beck, Alan ja Friedman, Batya (2009) "Robotic Pets in Human Lives: Implications for the Human–Animal Bond and for Human Relationships with Personified Technologies", *Journal of Social Issues* 65 (3), 545–567.

Melson, Gail F., Kahn Jr., Peter H., Beck, Alan, Friedman, Batya, Robert, Trafce, Garrett, Erik ja Gill, Brian T. (2009) "Children's Behavior Toward and Understanding of Robotic and Living Dogs", *Journal of Applied Developmental Psychology* 30 (2), 92–102.

Moutier, Sylvain, Angeard, Nathalie ja Houde, Olivier (2002) "Deductive Reasoning and Matching-Bias Inhibition Training: Evidence from a Debiasing Paradigm", *Thinking & Reasoning* 8 (3), 205–224.

Müller, Christian P. ja Schumann, Gunter (2011) "Drugs as Instruments: A New Famework for Non-Addictive Psychoactive Drug Use", *The Behavioral and Brain Sciences* 34 (6), 293–310.

Napier, Jaime L. ja Luguri, Jamie B. (2013) "Moral Mind-Sets: Abstract Thinking Increases a Preference for 'Individualizing' Over 'Binding' Moral Foundations", *Social Psychological and Personality Science* 4 (6), 754–759.

Nicholson, Nigel, Soane, Emma, Fenton-O'Creevy, Mark ja Willman, Paul (2005) "Personality and Domain-Specific Risk Taking", *Journal of Risk Research* 8 (2), 157–176.

Oldfield, Frank, Barnosky. Anthony D., Dearing, John, Fischer-Kowalski, Marina, McNeill, John, Steffen, Will ja Zalasiewicz, Jan (2014) "The Anthropocene Review: Its significance, Implications and the Rationale for a New Transdisciplinary Journal", *The Anthropocene Review* 1 (1), 3–7.

O'Neil, Cathy (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Palazzo, Guido, Krings, Franciska ja Hoffrage, Ulrich (2012) "Ethical Blindness", *Journal of Business Ethics* 109 (3), 323–338.

Palomäki, Jussi, Laakasuo, Michael ja Lappi, Otto (2012) "Ihmisen Emootiot, Päätöksenteko ja Rationaalisuus", *Ajatus,* 69, 91–120.

Preston, Jesse L. (2018) "The egocentric teleological bias: How self-serving morality shapes perceptions of intelligent design" teoksessa K. Gray ja J. Graham (toim.), *Atlas of moral psychology*. New York: The Guilford Press, 352–359.

Putt, Shelby S., Wijeakumar, Sobanawartiny, Franciscus, Robert G. ja Spencer, John P. (2017) "The Functional Brain Networks That Underlie Early Stone Age Tool Manufacture", *Nature Human Behaviour* 1 (6), 0102.

Rabin, Matthew ja Vayanos, Dimitri (2010) "The Gambler's and Hot-Hand Fallacies: Theory and Applications", *Review of Economic Studies* 77 (2), 730–778.

Riek, Laurel D., Rabinowitch, Tal-Chen, Chakrabarti, Bhismadev ja Robinson, Peter (2009) "How Anthropomorphism Affects Empathy Toward Robots", *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction,* 245–246.

Rode, Catrin, Cosmides, Leda, Hell, Wolfgang ja Tooby, John (1999) "When and Why Do People Avoid Unknown Probabilities in Decisions Under Uncertainty? Testing Some Predictions from Optimal Foraging Theory", *Cognition* 72 (3), 269–304.

Rothbart, Daniel ja Barlett, Tom (2008) "Rwandan radio broadcasts and Hutu/Tutsi positioning" teoksessa F. Moghaddam, R. Harré ja N. Lee (toim.), *Global conflict resolution through positioning analysis.* Springer Science & Business Media, 227–246.

Schäffner, Vanessa (2018) "Caught Up in Ethical Dilemmas: An Adapted Consequentialist Perspective on Self-Driving Vehicles", *Robophilosophy/TRANSOR*, 327–335.

Selby, Edwin C., Shaw, Emily J. ja Houtz, John C. (2005) "The Creative Personality", *Gifted Child Quarterly* 49 (4), 300–314.

Sharkey, Amanda ja Sharkey, Noel (2012) "Granny and the Robots: Ethical Issues in Robot Care for the Elderly", *Ethics and Information Technology* 14 (1), 27–40.

Shim, Jaeeun ja Arkin, Ronald C. (2013) "A Taxonomy of Robot Deception and Its Benefits in HRI", *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2328–2335.

Silva, Francisco J. ja Silva, Kathleen M. (2006) "Humans' Folk Physics Is Not Enough to Explain Variations in Their Tool-Using Behavior", *Psychonomic Bulletin & Review* 13 (4), 689–693.

Sober, Elliot ja Wilson, David Sloan (1999) *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Steinhoff, Uwe (2018) "The Secret to the Success of the Doctrine of Double Effect (and Related Principles): Biased Framing, Inadequate Methodology, and Clever Distractions", *The Journal of Ethics*, *22*(3-4), 235–263.

Tegmark, Max (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

Tononi, Giulio (2012) *PHI: A Voyage from the Brain to the Soul*. Pantheon Books.

Tooby, John ja Cosmides, Leda (2005) "Conceptual foundations of evolutionary psychology" teoksessa D. M. Buss (Toim.), *The Handbook of Evolutionary Psychology*. Hoboken: Wiley, 5–67.

van Leeuwen, Florian, Dukes, Amber, Tybur, Josbua M. ja Park, Justin H. (2017) "Disgust Sensitivity Relates to Moral Foundations Independent of Political Ideology", *Evolutionary Behavioral Sciences* 11 (1), 92–98.

Varoufakis, Yanis (2016) *And the Weak Suffer What They Must? Europe, Austerity and the Threat to Global Stability*. Random House.

Varoufakis, Yanis (2017) *Adults in the Room: My Battle with Europe's Deep Establishment*. Random House.

Visala, Aku (Ilmestyy 2021) "Moraalinen toimijuus ja ihmiskeskeisyyden dilemma tekoälyn maailmassa" teoksessa P. Raatikainen (toim.), *Tekoäly, ihminen ja yhteiskunta.* Helsinki: Gaudeamus.

Voci, Alberto (2006) "The Link Between Identification and In-Group Favouritism: Effects Of Threat To Social Identity And Trust-Related Emotions", *British Journal of Social Psychology*, 45 (2), 265–284.

Wachsmuth, Ipke (2018) "Robots Like Me: Challenges and Ethical Issues in Aged Care", *Frontiers in Psychology* 9, 432.

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

Ward, Adrian F., Olsen, Andrew S. ja Wegner, Daniel M. (2013) "The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead", *Psychological Science* 24 (8), 1437–1445.

Warwick, Kevin (2013) *Artificial Intelligence: The Basics.* Routledge.

Waytz, Adam, Epley, Nicholas ja Cacioppo, John T. (2010) "Social Cognition Unbound: Insights into Anthropomorphism and Dehumanization", *Current Directions in Psychological Science* 19 (1), 58–62.

Whitby, Blay (2008) "Sometimes It's Hard to Be a Robot: A Call For Action on the Ethics of Abusing Artificial Agents", *Interacting With Computers* 20 (3), 326–333.

Zeigler-Hill, Virgil, Noser, Amy, Roof, Courtney, Vonk, Jennifer, & Marcus, David K. (2015) "Spitefulness and Moral Values", *Personality and Individual Differences* 77, 86–90.