ROBOTIIKAN MORAALIPSYKOLOGIAN TUTKIMUS ON VÄLTTÄMÄTÖNTÄ

MICHAEL LAAKASUO JA JUSSI PALOMÄKI

Itsenäisiä päätöksiä tekevien robottien ja muiden tekoälyjen määrä lisääntyy valtavan nopeasti; tällaisia ovat esimerkiksi itseohjautuvat autot, hoitorobotit, robottipoliisit ja profilointialgoritmit. Kyseiset tekoälyt joutuvat enenevässä määrin tekemään myös moraalisia päätöksiä, jotka tavalla tai toisella liittyvät ihmisten hyvinvointiin (Bonnefon ym., 2016). Näin ollen moniin robotteihin pitää pian pystyä myös ohjelmoimaan ”moraalinen koodi” tai ohjenuora, jota noudattaa. Emme kuitenkaan vielä juuri lainkaan tiedä, minkälaista moraalia ihmiset haluavat tekoälyjen noudattavan tai mitkä tilannetekijät ylipäätään vaikuttavat ihmisten ja moraalisten robottien väliseen vuorovaikutukseen.

Ihmisen moraalinen kognitio on evoluutiohistorian aikana kehittynyt aivan toisenlaisessa ympäristössä kuin missä nyt elämme. Esi-isämme oppivat tekemään yhteistyötä toisten ihmisten kanssa, ymmärtämään heitä ja jakamaan yhdessä sekä resursseja että kokemuksia (Tooby ja Cosmides, 2005). Moraalikognitiomme on toisin sanoen satojen tuhansien vuosien ajan virittynyt reagoimaan muihin ihmisiin; mutta nyt se on pakotettu reagoimaan myös robotteihin ja keinoälyihin. Tilanne on ihmiskunnan evoluutiohistorian mittakaavassa poikkeuksellinen ja ainutlaatuinen. Ihmisille on esimerkiksi tyypillistä se, että he *antropomorfisoivat* (kuvittelevat ihmisen kaltaisiksi) keinotekoisia moraalisia toimijoita ja arvioivat niiden toimintaa omista inhimillisistä lähtökohdistaan (Duffy, 2003).

Kehityspsykologiassa ja moraalisen kognition tutkimuksessa robotit ja tekoälyt nähdään niin sanottuna uutena ontologisena kategoriana; ne pitäisi ymmärtää työkaluista, kasveista, ihmisistä ja muista eläimistä erillisenä, uutena olemassaolon muotona (Severson ja Carlson, 2010). Ihmisille on luontaisesti vaikeaa hahmottaa sitä, että sosiaalinen robotti, joka vain matkii älykästä ja tuntevaa oliota, ei oikeasti ole älykäs tai tunteva. Ihmiset esimerkiksi pahastuvat siitä, jos joku potkaisee robottikoiraa tai lyö puhuvaa ja ”söpöä” robottia (Melson ym., 2009). Tyypillisesti kirjallisuudessa tehdään jako *keinotekoisiin* ja *aitoihin* moraalisiin toimijoihin. Keinotekoiset moraaliset toimijat eivät ole tietoisia, ja niiden käyttäytyminen perustuu täysin ennalta ohjelmoituihin sääntöihin. Aidoilla moraalisilla toimijoilla puolestaan on sisäsyntyinen motivaatiorakenne ja tavoitteellisuus tehdä jotain aidosti hyvää tai pahaa. Robotit ja tekoälyt ja niihin kohdistuva ”väkivalta”, kuitenkin saavat meissä aikaan moraalireaktioita ja -tunteita. Emme kykene intuitiivisesti hahmottamaan robottien ja muiden tekoälyjen toiminnallisuutta oikealla tavalla, ja siksi saatamme tehdä virheitä niiden suunnittelussa ja niihin kohdistuvissa riskianalyyseissä.

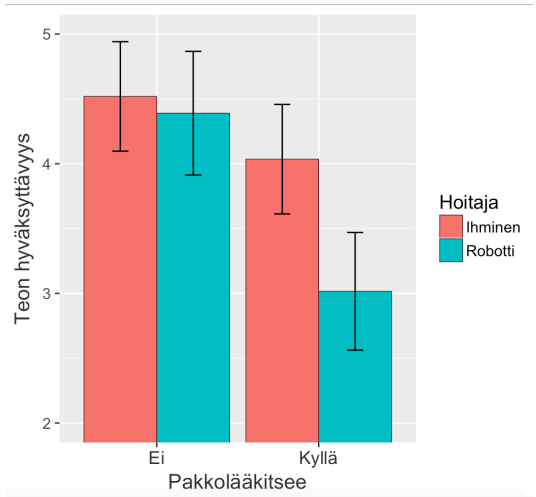
Myös monilla emootioilla ja tunteilla on oleellinen rooli ihmisten moraalisessa toiminnassa; näitä ovat esimerkiksi suuttumus, inho, kateus sekä oikeudenmukaisuuden ja reiluuden tunteet. Erityisesti inhon tunne näyttäisi olevan keskeisessä asemassa tekojen moraalista tuomittavuutta arvioitaessa: ihmiset paheksuvat tekoja, jotka tuntuvat inhottavilta (esim. insestiä tai oman maan lipun polttamista; Schnall ym., 2008). Suuttumus puolestaan motivoi ihmisten halua rangaista tai pyrkiä estämään tekoja, joita he pitävät tuomittavina. Moraalisesti paheksuttavien tekojen rankaiseminen on yhteisöjen koheesion ja toimivuuden kannal-

ta hyvin tärkeää, sillä rankaisu purkaa moraalisen suuttumuksen ja tilanne tietyllä tavalla normalisoituu; rankaisun jälkeen ”oikeus on tapahtunut” ja elämä voi jatkua.

Toistaiseksi juuri yhtäkään robottia tai muuta tekoälytoimijaa ei kuitenkaan ole suunniteltu ja rakennettu siten, että yllä käsiteltyjä ihmisten luontaisia tunnereaktioita otettaisiin huomioon. Esimerkiksi jos itseään ajava auto ajaa lapsen yli, se aiheuttaa vahvaa suuttumusta ja turhautumista – ja halun rangaista syyllistä. Tällaisessa tilanteessa ihmisten tunteet ja luontainen oikeustaju eivät todennäköisesti kuitenkaan tule tyydytetyksi, koska itseohjautuvan auton rankaiseminen on järjetöntä; eikä ole selvää, kehen muuhunkaan rankaisu voisi kohdistua. Toisaalta tiedetään, että esimerkiksi robottiprostituutio ja muistisairauksia parantavat aivoimplantit ovat tulevaisuudessa markkinoille tulevaa teknologiaa. Molemmat aiheuttavat omien alustavien tutkimustemme mukaan ihmisissä inhoreaktioita; jos uusi (ja potentiaalisesti nopeasti leviävä) teknologia aiheuttaa yleisesti vahvaa inhoa ja paheksuntaa, näitä reaktioita ja niiden syitä tulee tutkia ennen kuin teknologia on yleisesti saatavilla.

Ainakin seuraavat kysymykset tulee ottaa vakavasti huomioon robotiikan ja tekoälyjen nopeassa kehityksessä: 1) Miten ihmiset suhtautuvat uuteen (moraaliseen) teknologiaan, joka ei välttämättä kunnioita ihmisten omia pyrkimyksiä ja vapaata tahtoa? 2) Miten ihmiset reagoivat tilanteissa, joissa tekoäly vahingossa tai tarkoituksellisesti tekee ihmisen hyvinvointiin vaikuttavia ratkaisuja? 3) Miten vastuun koetaan jakautuvan tilanteissa, joissa tekoälyjen tekemät päätökset johtavat mahdollisiin onnettomuuksiin? Näitä kysymyksiä tulisi tarkastella demokraattisten rakenteiden näkökulmasta: mitä enemmän moraalisia, ihmisen hyvinvointiin vaikuttavia päätöksiä automatisoidaan, sitä tärkeämpää on, että ihmiset itse pääsevät vaikuttamaan näiden automaattisten koneiden ”moraaliseen koodin” suunnittelussa.

Tuoreessa julkaisemattomassa tutkimuksessamme arvioimme kokeellisesti ihmisten suhtautumista hoitorobottien (verrattuna ihmishoitajien) tekemiin päätöksiin kuvitellussa tilanteessa, jossa sairaalan ylilääkäri antaa käskyn pakkolääkitä vastentahtoinen potilas. Kuva 1 selventää toistuvasti



Kuva 1. Ihmishoitajan tai hoitorobotin tekemän pakkolääkitsemispäätöksen hyväksyttävyys (asteikolla 1–7; virhemarginaalit ovat 95 %:n luottamusvälejä).

havaittavaa kaavaa: ihmishoitajan päätökset joko pakkolääkitä potilas (noudattaa ylilääkärin ohjeita) tai jättää hänet pakkolääkitsemättä (uhmata ylilääkärin ohjeita) koetaan yhtä hyväksyttäviksi; mutta jos pakkolääkitsemisen toteuttaa hoitorobotti, sen toimintaa paheksutaan selvästi enemmän (tai hyväksytään vähemmän, kuten kuvassa 1). Jatkotutkimuksissa olemme havainneet, että kuvassa 1 esitetty kaava korostuu, mikäli pakkolääkitseminen (tai lääkitsemättä jättäminen) johtaa potilaan kuolemaan; mutta hoitorobotin teknisellä luotettavuudella tai toimintavarmuudella ei ole merkitystä sen päätösten arvioinnissa. Tämä viittaa siihen, että sairaaloiden automatisoinnissa saatetaan hyväksyä käyttöön keskinkertaisia ratkaisuja ja olla niihin tyytyväisiä, vaikka parempaakin olisi todennäköisesti tarjolla.

Ihmisten toiminta emootioita herättävissä moraalipsykologisissa dilemmatehtävissä välittää meille tietoa luontaisen moraalitajumme rakentumisesta ja rajoista. Tutkimuksissa voidaan verrata tekoälyjen ja ihmisten tekemien moraalisten päätösten arvioita, ja sitä kautta paljastaa epäjohdonmukaisuuksia ihmisten ajattelussa: vaikka ihminen

ja robotti tekisivät saman moraalisen päätöksen, kyseisen päätöksen hyväksyttävyys koetaan täysin eri tavalla. Tämä on oleellista tietoa myös lakien säätäjille ja hallintohenkilöille, jotka päättävät robottien hankinnasta mm. vanhusten- ja sairaanhoidossa. Moraalipsykologia on viime aikoina edistynyt merkittävästi, mutta tieteenala ei kuitenkaan kehity tällä hetkellä riittävän nopeasti tuodakseen selkeyttä teknologian kehityksen aikaansaamiin uusiin pulmatilanteisiin. Tästä johtuen robotiikan moraalipsykologian tutkiminen on sekä hedelmällistä että välttämätöntä.

Jo nyt algoritmit ja tekoälyt käyvät läpi erittäin laajoja tietokantoja ja tekevät suosituksia vakuutusyhtiöille, viranomaisille ja poliittisille toimijoille. Esimerkiksi erilaisilla ”web-scraping” -teknologiaan pohjaavilla algoritmeilla voidaan kerätä yhden henkilön tietoja useista eri lähteistä ja luoda kattava profiili tästä henkilöstä häneltä itseltään mitään kysymättä. Onko moraalisesti oikein ohjelmoida sellaisia ohjelmia, jotka tiivistävät kattavasti koko elämänhistoriamme internetin käyttömme perusteella? On täysin mahdollista, että tulevaisuudessa googlailuhistoriamme takia meiltä esimerkiksi evätään henkivakuutus.

Tammikuussa 2016 julkistettiin Yhdysvalloissa ensimmäiset katuja partioivat robottipoliisit, ja on selvinnyt, että suuriin aineistoihin perustuvia profilointialgoritmeja käytetään ihmisten ehdonalaislupien myöntämisessä. Dubaissa robottipoliisit otettiin käyttöön niin ikään vuonna 2016, ja lokakuussa Saudi-Arabia antoi kansalaisuuden androidille[1]. Myös Suomen eduskunnassa järjestettiin keskustelutilaisuus yhteiskuntamme robotisoimisesta kesäkuussa 2016. Kaikki nämä seikat viittaavat siihen, että robotiikan moraalipsykologian tutkimus on sekä ajankohtaista että välttämätöntä; sille on selkeä paikka tieteiden ja yhteiskunnallisten tarpeiden kentällä.

Lähteet

Bonnefon, J. F., Shariff, A. ja Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573–1576.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems*, *42*(3), 177–190.

Melson, G. F., Kahn Jr, P. H., Beck, A. ja Friedman, B. (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.

Severson, R. L. ja Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, *23*(8), 1099–1103.

Schnall, S., Haidt, J., Clore, G. L. ja Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and social psychology bulletin*, *34*(8), 1096–1109.

Tooby, J. ja Cosmides, L. (2005). Conceptual foundations of evolutionary psychology, teoksessa *The Handbook of evolutionary psychology* (ed. Buss, M.), 5–67. John Wiley ja Sons.

1 http://www.independent.co.uk/life-style/gadgets-and-tech/news/saudi-arabia-robot-sophia-citizenship-android-riyadh-citizen-passport-future-a8021601.html

Michael Laakasuo ja Jussi Palomäki ovat kognitiotieteen tutkijatohtoreita sekä työskentelevät Moralities of Intelligent Machines -tutkimusryhmässä (www.moim.fi). Laakasuo on ryhmän vastaava tutkija.

TUKEA TEKOÄLYHANKKEILLE

Erilaisten datalähtöisten menetelmien merkitys tutkimuksessa, hallinnossa ja teollisuudessa kasvaa jatkuvasti. Tekoälymenetelmien taustalla ovat koneoppimisen, hahmontunnistuksen, tilastotieteen, tiedonlouhinnan ja tietokantatekniikoiden laskennallisten ja ohjelmistoteknisten menetelmien merkittävät edistysaskeleet sekä laskentatehon nopea kasvu. Näillä uusilla menetelmillä on laajoja sovelluksia myös tieteenteossa.

Suomen Akatemia myönsi viime vuoden lopussa ohjelmallisesta rahoituksesta yhteensä yli 13 miljoonaa euroa tekoälytutkimukseen. ICT2023-ohjelman tekoälyhankkeille myönnettiin reilut kuusi miljoonaa euroa ja ”Tekoälyn uudet sovellukset fysikaalisten tieteiden ja tekniikan tutkimuksessa” (AIPSE) -akatemiaohjelman tutkimushankkeille yhteensä seitsemän miljoonaa euroa. ICT2023-hankkeissa parannetaan biolääketiedealan koneoppimismenetelmiä ja kehitetään ratkaisuja likimääräisten algoritmien päättelyn laadun arviointiin. AIPSE-hankkeissa kehitetään laskennallisia menetelmiä ruoan tuotannon tehostamiseen ja haetaan läpimurtoa nanomateriaalien laskennallisessa tutkimuksessa.