Aku Visala

Ihmiskäsitykset tekoälyn aikakaudella

Pyrin tässä paperissa perustelemaan väitettä, jonka mukaan tekoälyn tutkimuksen ja käytön ihmiskäsitykseen liittyvät vaikutukset ja sidokset tulisi ottaa etiikassa ja politiikassa huomioon. Tarkoitus on nostaa esiin kysymyksiä ja huolenaiheita, jotka eivät ole luonteeltaan pelkästään taloudellisia tai yksittäisiin moraalisiin dilemmoihin liittyviä.

Miksi ihmiskäsityksillä on merkitystä?

Yhdysvaltain avaruushallinto päätti vuonna 2015 rahoittaa usealla miljoonalla dollarilla kolmivuotista projektia, jonka tutkimuskohteena ovat astrobiologian yhteiskunnalliset ja uskonnolliset vaikutukset. Tässä projektissa tarkastellaan erilaisia poliittisia, moraalisia, teologisia ja muita yhteiskunnallisia vaikutuksia, joita planeettamme ulkopuolisen älykkään elämän etsintä (tai sen löytyminen) voi laukaista. NASA pitää tällaista työtä tarpeellisena erityisesti siksi, että sen oma olemassaolo riippuu ihmisten verorahoista ja näin ollen siitä, miten ihmiset siihen suhtautuvat. Projektin ytimessä on kuitenkin tärkeä huomio: maan ulkopuolisen älykkyyden etsinnällä ja mahdollisella löytämisellä on valtavat vaikutukset siihen, miten ihmiskunta kokonaisuutena näkee itsensä, oman arvonsa ja päämääränsä kosmoksessa. Tämä kysymys ei ole pelkästään tieteellinen tai eettinen, vaan maailmankatsomuksellinen, uskonnollinen ja poliittinen.

Samankaltaista päättelyä voidaan soveltaa planeetallamme harjoitettavaan tekoälyn tutkimukseen ja mahdolliseen kehitykseen. Sen lisäksi, että ihmiskäsitykset ohjaavat joiltakin osin sitä, millaista tekoälyä pyritään kehittämään ja miten tässä edetään, on tekoälyn tutkimuksella ja tulevalla arkielämää muuttavalla käytöllä todennäköisesti vaikutus siihen, miten ihmiset ylipäänsä näkevät itsensä. Tutkimuskirjallisuudessa puhutaan erilaisista strategisista uhkista ja haasteista, joita tekoälyn kehitys voi tuoda mukanaan. Esimerkiksi Nick Bostromin kirjassa *Superintelligence* (2014) ei kuitenkaan edes mainita ihmiskäsitykseen liittyviä mahdollisia uhkia ja mahdollisuuksia. Millaisia vaikutuksia ihmiskäsitykseemme on esimerkiksi sillä, että kapea tekoäly ylittää ihmisen kyvyt jollakin tietyllä alueella, yleinen tekoäly saavuttaa ihmisen kapasiteetin monella alueella tai superälykkyys syntyy? Miltä osin politiikka muuttuu? Miten erilaiset katsomukselliset yhteisöt suhtautuvat tähän?

Tällaiset kysymykset ovat olennaisia tekoälyn etiikan kannalta useista syistä. Ensinnäkin ihmiskäsityksillä on keskeinen rooli etiikan ja politiikan taustatekijöinä. Moniarvoisessa yhteiskunnassa myös ihmiskäsityksissä on jonkin verran variaatiota. Filosofisiin, uskonnollisiin ja poliittisiin perinteisiin liittyy erilaisia oletuksia siitä, millaisia ihmiset ovat, mikä on heille hyvää ja millainen heidän psykologiansa on. Tämä variaatio näkyy esimerkiksi puolueiden ja uskonnollisten yhteisöjen eettisissä kannoissa. Muutokset ihmiskäsityksissä voivat aiheuttaa suuriakin muutoksia politiikassa ja etiikassa. Tämä nähdään jo nyt bioetiikan alueella. Ihmiskäsitykset vaikuttavat merkittävästi siihen, miten eri ryhmät ja yhteisöt suhtautuvat uuteen teknologiaan ja niiden kehittämiseen yhteiskunnan avustuksella. Yhteiskunnan digitalisoituessa olisi siis tärkeää ottaa huomioon ihmiskäsitykset, koska ne määrittävät merkittävästi ihmisten asenteita tätä kehitystä kohtaan. Lisäksi erimielisyydet ja jännitteet näiden teemojen äärellä lisäävät rauhattomuutta ja epävakautta.

Toinen syy, miksi ihmiskäsityksistä on tärkeä keskustella, koskee tekoälyteknologian ”ihmismäistä” luonnetta. Teknologia ylipäänsä ei ole arvoneutraalia, koska sen tuotanto ja käyttö vaikuttaa ihmissuhteiden luonteeseen, yhteisöihin ja yksilöihin tavoilla, jotka ovat joskus ennakoimattomia. Teknologiaa suunnitellaan aina johonkin tarkoitukseen, joten se antaa konkreettisen muodon suunnittelijan päämäärille ja tätä kautta myös arvoille. Tekoälyteknologiat ovat erityisen haastavia juuri näiltä osin, koska niissä pyritään ainakin joiltakin osin mallintamaan ja käsittelemään ihmisyyden ydinalueita. Näitä ovat esimerkiksi empatia, erilaiset älykkyyden muodot, tunteet, ihmissuhteet, vuorovaikutus sekä moraalinen käyttäytyminen. Näillä alueilla käytettävät teknologiat voivat puolestaan vaikuttaa siihen, kuinka ihmiset nämä seikat näkevät ja millaisia asioita he pitävät arvokkaina. Ihmiskäsitykset eivät siis pelkästään ohjaa tekoälyn kehittymistä, vaan tekoälyn käyttö ja kehitys vaikuttaa puolestaan ihmiskäsityksiimme. Emme esimerkiksi vuonna 2000 osanneet arvata millaisen muutoksen inhimillinen vuorovaikutus kokee sosiaalisen median tullessa osaksi arkielämäämme. Tekoälyn ja robotiikan kohdalla nämä vaikutukset liittyvät vuorovaikutuksen lisäksi esimerkiksi rakkaussuhteisiin, hoivaan sekä ihmisten kohtaamiseen.

Kolmas syy tarkastella ihmiskäsityksiä on niiden rooli siinä, miten ja millaisia tekoälyjärjestelmiä kehitetään. Myös näiden järjestelmien kehittäjät ovat ihmisiä, joilla on arvonsa ja ihmiskäsityksensä. Jos suunnittelijoiden mielessä ihmisyyden ideaali on utilitaristinen insinööri, on tällä varmasti vaikutus siihen, millaista älykkyyden muotoa pidetään tavoittelemisen arvoisena. Jos taas suunnittelija tunnistaa omat ajatusvinoumansa, kykenee hän (ehkä) välttämään näiden siirtämisen algoritmeihinsa. Tässä vain muutama esimerkki.

Neljänneksi on tärkeä huomata, ettei ihmiskäsityksissä ole kyse vain ja ainoastaan yksityisten ihmisten ja maailmankatsomuksellisten ryhmien näkemyksistä. Ihmiskäsitysten keskeisiä piirteitä on kirjattu kansallisiin ja kansainvälisiin lakeihin. Tärkeä osa eurooppalaista ja myös suomalaista arvopohjaa on tietynlainen käsitys siitä, millaisia ihmiset ovat ja mikä on heille hyvää. Esimerkiksi ihmisarvo, moraalinen autonomia, erilaiset vapaudet ja velvollisuudet sekä perusoikeudet implikoivat jo tällaisen ihmiskäsityksen. Nämä toimivat tulevan lainsäädännön ohjenuorina sekä ohjaavat alempien tasojen periaatteiden tulkintaa.

Mitä ihmiskäsitykseen kuuluu?

”Ihmiskäsitys” on joiltakin osin epäselvä termi. Viittaan sillä tässä yhteydessä joukkoon erilaisia seikkoja. Länsimaisen filosofian ja teologian perinteessä tässä yhteydessä käsitellään esimerkiksi seuraavia kysymyksiä:

1. Millaisia olioita me olemme? Onko meillä esimerkiksi sielu ja millainen se on luonteeltaan? Edellyttääkö tietoisuus sielua?
2. Millainen ihmisen psykologia on? Millaisia psykologisia kykyjä meillä on?
3. Ovatko ihmiset erityisiä verrattuna muihin eläimiin? Mitkä voisivat olla näitä erityispiirteitä? Onko ihmisellä luonto tai olemus?
4. Onko ihmisellä vapaa tahto? Ovatko hänen tekonsa hänestä itsestään kiinni vai onko hän esimerkiksi kohtalon tai luonnon ohjauksessa?
5. Ovatko ihmiset luonnostaan hyviä vai pahoja, itsekkäitä vai epäitsekkäitä?

6. Mistä olemme tulleet ja mihin olemme menossa? Onko olemassaololollamme jokin objektiivinen tarkoitus?
7. Millainen yhteiskunta on kaltaisillemme olioille sopiva?
8. Millainen on esikuvallinen ihminen? Millainen ihmisen ideaali meillä on? Mitä viisaus on?

Psykologisissa tutkimuksissa on lisäksi tullut ilmi, että ihmisillä on ainakin joiltakin osin jaettu, intuitiivinen ihmiskäsitys. Tähän kuuluu esimerkiksi oletus, että kaikkia ihmisiä yhdistää jonkinlainen ”olemus” tai luonto, joka on viime kädessä kausaalisessa vastuussa ihmisiä yhdistävistä näkyvistä piirteistä ja taipumuksista. Ihmisillä on vahva taipumus liittää tämä olemus biologiaan ja siihen sisäänrakennettuihin päämääriin. Yksilöitä ja ryhmiä erottavia piirteitä selitetään puolestaan usein kulttuurieroilla. Toiseksi on jonkin verran näyttöä siitä, että ihmisillä on taipumus olettaa jonkinlainen minuus tai sielu, joka tekee kustakin yksilöstä erityisen ja antaa tälle erityisaseman verrattuna muihin eläimiin. Tämä sielu on myös tavalla tai toisella ihmisälyn ja ihmisarvon erityinen asuinsija. Kolmanneksi ajatus siitä, että kaikilla ihmisillä on luontonsa ja olemuksensa puolesta jonkinlaisia objektiivisia päämääriä, on varsin tavallinen.

On myös hyvä erottaa toisistaan maallikkojen enemmän tai vähemmän tietoiset ihmiskäsitykset, jotka nousevat monista erilaisista lähteistä, sekä eri alojen tutkijoiden, tutkimukseen perustuvat näkemykset. Tekoälyn kehitys ja käyttö vaikuttaa molempiin.

Ihmiskäsitysten ja etiikan rajanveto on toisinaan vaikeaa. Kysymykset siitä, mikä on hyvää ja millainen yhteiskunta on ihmiselle hyväksi, kuuluvat selkeästi myös etiikan alaan. Toisaalta taas esimerkiksi kysymykset ihmisen erityisyydestä ja ihmisluonnosta eivät ole selkeästi eettisiä kysymyksiä. Ihmiskäsityksen tasolla tehdyt oletukset vaikuttavat siihen, millä tavalla etiikkaa ja politiikkaa lähdetään tekemään ja millaiset ”korkean tason” eettisen periaatteet valitaan. Tämä ihmiskäsitysten ja etiikan kietoutuminen nähdään jo nyt bioetiikassa. Jatkossa se nähdään yhä vahvemmin tekoälyn etiikassa. Vaikka ihmiskäsityksillä onkin eettinen ulottuvuutensa, ovat ne kuitenkin periaatteessa erillisiä normatiivisesta etiikasta. Ihmiskäsitys muodostaa erään etiikan ja politiikan taustatekijän: normatiiviseen etiikkaan kuuluu joukko ”keskitason” periaatteita ja normeja, joita voidaan perustella myös toisistaan poikkeavien ihmiskäsitysten avulla.

Joitakin tekoälyn mahdollisia vaikutuksia ihmiskäsityksiin

Mainitsen lopuksi muutamia erityisteemoja ja kysymyksiä, jotka koskevat ihmiskäsitystä ja tekoälyä. Nämä teemat ovat vain pieni ja pintapuolinen otos kaikista mahdollisista kysymyksistä.

Voi olla, että lisääntyvä vuorovaikutus tekoälyjärjestelmien kanssa vaikuttaa siihen, millaisen moraalisen statuksen annamme toisillemme ja miten suhtaudumme toisiimme. Jotkut filosofit (esim. Tuomas Akvinolainen ja Immanuel Kant) ovat esittäneet, että sillä, miten kohtelemme muita eläimiä, on vaikutus moraalipsykologiaamme. Muita eläimiä on tämän argumentin mukaan syytä kohdella hyvin (vaikka näillä ei olisikaan tietoisuutta), koska julma kohtelu tukee taipumustamme kohdella toisia ihmisiä julmasti. Jos argumentti toimii, se voidaan laajentaa koskemaan myös tekoälyjärjestelmiä. Voiko käydä niin, että yhä enemmän sosiaalisesti ja ”ihmisenkaltaisesti” käyttäytyvän tekoälyjärjestelmän kohtelu vaikuttaa moraalisiin taipumuksiimme negatiivisesti?

Vastaus riippuu siitä, millä tavalla tosiasiassa kohtelemme näitä järjestelmiä. Ongelmia voi syntyä siitä, että pidätämme inhimillisen kohtelun keino-olioilta, jotka ovat monilta osin ihmisen kaltaisia. Tätä teemaa on käsitelty laajasti fiktiossa, esimerkiksi 2017 ilmestyneessä *Blade Runner 2049* elokuvassa. Elokuvassa oikeat ihmiset ovat tottuneet kohtelemaan ihmisenkaltaisia replikantteja huonosti, mikä on puolestaan tehnyt heistä itsestään kovapintaisia ja piittaamattomia kaiken, myös toistensa, kärsimyksen suhteen. Ihmisestä itsestään voi siis tulla tekoälyjen yleistyessä koneen kaltainen.

Toisaalta taas ongelmia voi syntyä siitäkin, että ulotamme inhimillisen kohtelun koskemaan olioita, jotka ovat hyvin kaukana ihmisen (erityisesti) moraalisista ja sosiaalisista kyvyistä. Tässä yhteydessä ei tarvitse tukeutua vain fiktiosta haalittuihin esimerkkeihin. Olemme jo nyt tilanteessa, missä jotkut ihmiset ottavat inhimillisen (ja intiimin) kohtelun piiriin erilaisia keino-olioita. Esimerkiksi jotkut ihmiset ovat halunneet mennä naimisiin robottiensa kanssa. Monet robotit suunnitellaan tietoisesti niin, että ne herättävät vahvoja moraalisia ja sosiaalisia reaktioita ihmisissä, vaikka näillä olioilla ei olekaan kykyä syvään vuorovaikutukseen, rakkauteen ja autonomiaan. Tällöin olemme vaarassa trivialisoida inhimillisen kohtelun ja sen edellytykset. Moraalisesti latautuneet, monimutkaiset ihmissuhteet voidaan ymmärtää yhä kapeammin. Robotin ja omistajan suhde on hyvin kaukana kahden ihmisen välisestä rakkaussuhteesta. Jos tätä kuitenkin aletaan pitää rakkaussuhteena sanan vahvassa merkityksessä, käsitys rakkaudesta kapenee. Rakkaudesta, joka on monimutkainen, uhrautumista, tunteita ja ruumiillista yhteyttä edellyttävä ja velvollisuuksia luova suhde, tulee lähinnä nimitys seksuaaliselle tyydytykselle ja vain yhden osapuolen tarpeiden täyttymiselle. Samankaltaista kehitystä voi tapahtua myös muiden, moraalisesti latautuneiden suhteiden, kuten hoivan ja huolenpidon suhteen. Vaikka tällaiset kysymykset ovatkin luonteeltaan enemmän eettisiä ja filosofisia, on empirialla myös oma roolinsa. Ihmisten ja robottien interaktiota ja sen vaikutuksia voidaan tutkia psykologian keinoin ja tällaista tutkimusta jo tehdäänkin.

Edellinen uhkakuva voidaan esittää myös hieman toisella tavalla. Filosofit Arto Laitinen ja Antti Kauppinen puolustivat esityksessään näkemystä, jonka mukaan moraalisesti relevantit päätökset tulee jatkossakin pitää ihmisten käsissä, koska tekoälyllä ei ole vielä pitkään aikaan kykyä olla vastuussa teoistaan. Tekoäly ei siis voi olla moraalinen toimija eikä oikeushenkilö nykyisessä muodossaan. Tätä väitettä voitaisiin perustella myös kiinnittämällä huomiota mahdollisiin negatiivisiin seurauksiin, joita keino-olioiden vastuussa pitämisellä ja yhä laajemmalla käytöllä voi olla. Kyky toimia moraalisesti korkealla tasolla edellyttää pääsyä erilaisten tilanteiden moraalisesti relevantteihin tosiseikkoihin (esim. millaiset asiat aiheuttavat kärsimystä, millaista apua ihmiset tarvitsevat, miten henkilö tunnistaa omat negatiiviset tunteensa ja kykenee myötätuntoon). Kyky tunnistaa tällaisia tosiseikkoja ei ole mitenkään automaattinen, vaan edellyttää harjaantumista. Kykyjen harjaantuminen riippuu kuitenkin varsin paljon kontekstista. On esimerkiksi paljon helpompi halveksua ja pilkata toista henkilöä sosiaalisessa mediassa kuin kahvipöydässä samaisen henkilön kanssa. Sosiaalinen media on muuttanut vuorovaikutuksen kontekstia ja sitä kautta vaikuttaa psykologiaamme. Tekoälyn yhä lisääntyvä läsnäolo arkielämässä voi tehdä samoin vielä suuremmalla mittakaavalla.

Kuvitellaan lähitulevaisuus, jossa tekoälyjärjestelmät ovat yhä enemmän läsnä ihmisten elämässä, esimerkiksi eri palvelualoilla, ja ne kykenevät yhä laadukkaampaan inhimilliseen vuorovaikutukseen. Tästä seuraa se, että joissakin rajatuissa tilanteissa, esimerkiksi kun soitan puhelimella pankkiin hoitamaan asioitani, on vaikea erottaa, onko vastapuolella ihminen vai kone.

Nämä järjestelmät eivät kuitenkaan kykene koko inhimillisen vuorovaikutuksen kirjoon eivätkä myöskään vastuuseen. Tällä voi olla useita seurauksia. Ihmiset voivat tottua kohtelemaan yhä ”inhimillisempiä” järjestelmiä välinpitämättömästi, koska niillä ei kuitenkaan ole ihmiskollegojensa vastuuta ja valtaa. Ihmiset voivat tulla myös yhä epäluuloisemmiksi myötätunnon osoituksia kohtaan, koska he herkistyvät niille seikoille, jotka paljastavat epäautenttisen, simuloidun, myötätunnon.

Tekoälyjärjestelmien kohtelu liittyy vastuuseen ja vapaaseen tahtoon myös toisella tavalla. Tutkimusten mukaan ihmisten arkikäsitys vapaudesta ja vastuusta edellyttää, että vastuullisella toimijalla on kyky tunnistaa asiaankuuluvat moraaliset periaatteet ja ohjata tekojaan oman tahtonsa mukaisesti. Otamme toisten ihmisten ja omat (tahdonalaiset) tekomme merkkeinä siitä, millaisia me olemme, ja kohtelemme itseämme ja muita sen mukaisesti. Joku on esimerkiksi luotettava, toinen epäluotettava; joku taas rohkea ja toinen pelkuri. Ihminen vetää tällaisia johtopäätöksiä omista ja toisten teoistaan pääsääntöisesti automaattisesti. Miten tällaisille automaattisille päättelyille käy tilanteessa, jossa tekoälyjärjestelmät toimivat yhä monipuolisemmin, mutta vailla minkäänlaista ”tahtoa” tai luonnetta?

Vapaa tahto ja vastuu ovat olennainen osa ihmiskäsitystä myös niiltä osin, että niitä koskevat oletukset ja uskomukset vaikuttavat merkittävästi ihmisten käytökseen. Tutkimuksissa on osoitettu lukuisia yhteyksiä vapaata tahtoa koskevien uskomusten ja esimerkiksi aloitekyvyn, luovuuden, sosiaalisen riskinoton ja rehellisyyden suhteen. Mitä vahvemmin ihmiset uskovat vapaaseen tahtoon, sitä enemmän heillä esiintyy edellisiä käyttäytymistaipumuksia ja asenteita. On siis ainakin jotakin käytännöllisiä syitä pitää yllä ja tukea ihmisten uskoa vapaaseen tahtoon. Seuraako tekoälyn yleistymisestä taipumus ajatella, että ihminenkin on vain tietyllä tavalla ohjelmoitu kone, mikä puolestaan näyttäisi johtavan tahdonvapauden epäuskottavuuteen? Vai seuraako siitä se, että ihmiset laajentavat vapautta ja vastuuta tavalla, joka voisi mahdollistaa myös tekoälyjen pitämisen vapaina toimijoina? Kummassakin skenaariossa on omat ongelmansa.

Kuten edellä kävi ilmi, ihmisillä on taipumus olettaa jonkinlainen ihmisen erityisyys sekä moraalisessa että metafyysisessä mielessä. Mitä tapahtuu tälle oletukselle, kun ihmisen ajattelua ja käyttäytymistä mallintavat järjestelmät lisääntyvät ja ihmiset kohtaavat niitä yhä useammin arkielämässään? Tämä voi edesauttaa kehitystä, jossa intuitiivinen käsitys ihmisen erityisyydestä alkaa rapautua. Tämä voi toki olla hyväkin asia: se voi esimerkiksi edesauttaa eläinten parempaa kohtelua. Seuraukset voivat kuitenkin olla myös vähemmän toivottuja: jos ihmisen erityisyys rapautuu, saattaa vaikkapa ajatus ihmisarvosta olla yhä vaikeampi perustella.

Voidaan myös kysyä, mitä tarkoitamme ”inhimillisellä ja ihmisarvoisella kohtelulla ja ihmisarvoisella kohtelulla”. Millaisia tekijöitä tähän kohteluun liittyy? Näitä ovat varmaankin ihmisten autonomian kunnioittaminen, ihmisarvo, vastuussa pitäminen sekä erilaiset vapaudet ja velvollisuudet. Voi hyvinkin olla niin, että tulevaisuudessa jotkut tekoälyjärjestelmät saavuttavat ihmisen kyvyt ja ylittävät ne alueilla, joissa ne nykyisin vielä ovat varsin kehittymättömiä (moraali, sosiaalinen älykkyys, tunteet, jne.). Tällöin kysymme, missä vaiheessa nämä järjestelmät ansaitsevat ”inhimillistä kohtelua”, mukaan lukien vastuussa pitämistä. Tähän ei voida vastata operoimalla pelkästään normatiivisen etiikan ”keskitason” periaatteiden alueella, vaan tarvitaan laajempaa keskustelua ihmiskäsityksistä.

Ihmiskäsitykset ovat keskeisessä roolissa, kun kehitämme pitkän tähtäimen visioita tulevaisuudesta ja ihmisen kohtalosta. Nämä eivät välttämättä kuulu itse selonteon alaan mutta on kuitenkin hyvä mainita ne tässä. Kiinnitin huomiota eri verkostojen yhteisessä seminaarissa siihen, että laajamittaiset narratiivit ihmisen tulevaisuudesta esitetään ainakin joiltakin osin uskonnollista terminologiaa käyttäen. Ihmiskäsityksillä on tärkeä osa tällaisissa narratiiveissa, koska niissä kuvataan ihmisen päämääriä, tavoitteita ja elämän merkitystä. Kuten tunnettua, muutamat transhumanistit puolustavat ja pyrkivät edesauttamaan tietynkaltaista teknologista kehitystä, jonka avulla ihmisen biologian rajat pyritään ylittämään. Tekoäly ja superälykkyys ovat osa tätä visiota: ne ratkaisevat ihmiskuntaa vaivanneet ongelmat ja ainakin osa näistä ongelmista (esimerkiksi kuolema ja resurssien puute) johtuu biologisen olemassaolomme ennakkoehdoista. Ainakin joidenkin transhumanistien mukaan meillä on moraalinen velvollisuus ylittää nämä rajat ja ennakkoehdot. Tekoäly kykenee siis pelastamaan ihmiskunnan. Transhumanismin kritiikkiä motivoi myös tietynlainen ihmiskäsitys, jossa ihmisen biologialla ja sen reunaehdoilla on keskeisempi rooli. Tällaisissa väittelyissä keskustellaan lopulta siitä, millaisia ihmisiä haluamme olla ja miten tuohon tilaan päästään.