TEKOÄLYN EETTISET RISKIT JA VASTUULLINEN SUUNNITTELU

Tekoälyteknologian kehitys on ollut viime vuosina räjähdysmäistä; mullistavalta tuntuvia läpimurtoja tapahtuu lähes kuukausittain. Uusia tekoälysovelluksia ilmestyy työpaikoille, kouluihin, sairaaloihin, valtionhallintoon, viihdeteollisuuteen, kaikkialle. Nopeuden lisäksi yhteiskunnallisten ja taloudellisten vaikutusten määrä on valtava. Puhutaan jopa neljännestä teollisesta vallankumouksesta ja yhteiskunnallisten muutosten odotetaan olevan samaa luokkaa kuin höyrykoneiden, sähkön ja digitalisaation olivat. Samalla monia epäilyttää muutoksen nopeus ja tekoälyn kyky korvata ihmistä ajattelua vaativissa toimissa. Jopa ns. digijättien, suurten tietotekniikkayritysten sisältä on kuulunut suorastaan pelokkaita puheenvuoroja.

Tekoälyteknologiaa ei kuitenkaan pidä mystifioida sen enempää kuin mitään muutakaan edistyksellistä teknologiaa. Emme ole rakentamassa ”jumalakoneita” tai ”singulariteettia”. Tekoälysovelluksiin liittyy silti monia vaikeita eettisiä kysymyksiä, joista osa on vanhoja, osa täysin uusia. Tämän artikkelin tarkoitus on esitellä joitakin tällaisia ongelmia ja kaksi erilaista tapaa yrittää elää näiden ongelmien kanssa: kone-etiikka eli eettisten normien sisällyttäminen tekoälyn toimintaan sekä arvosensitiivinen suunnittelu. Ensin kuitenkin lyhyt luonnehdinta siitä, mistä tekoälyssä on kyse.

Mitä on tekoäly ja mihin sitä käytetään?

”Tekoälylle” ei ole kovin hyvää määritelmää. Yksi yleinen tapa kiertää määritelmä on sanoa, että tekoälyä ovat ne ohjelmistot, jotka kykenevät suorittamaan tehtäviä, joihin ihminen tarvitsisi älykkyyttä – miten tämä sitten määritelläänkään. Tekoäly ei kopioi ihmisen ajattelun prosesseja ja ohjelmistot toimivat melko mekaanisesti, mutta ne mahdollistavat älykkyyttä vaativien toimintojen automatisoimisen. Tärkeä osa tekoälyteknologiaa on koneoppiminen ja erityisesti niin sanottu syväoppiminen. Siinä algoritmi muuttaa toimintaansa palautteen perusteella ja pystyy luomaan monimutkaisiakin assosiaatioita ja luokitteluja. Opittuja assosiaatioita voidaan käyttää erilaisissa data-analyyseissä, päätöksentekoon ja jäljittelemään esimerkiksi kielellisten ilmaisujen tuottamista. Assosiaatiot eivät kuitenkaan ole varsinaisesti päättelyiden tekemistä, ja oppiminen on luonteeltaan behavioristista.

Tekoälytyökaluja on kehitetty esimerkiksi datan analyysiin lääketieteessä, poliisitoimintaan ja tiedusteluun, työhönottoon, korvaamaan byrokratiaa virastoissa ja niin edelleen. Kielimallit tarjoavat apua tekstin muokkaamiseen ja raakatekstin luomiseen. Tällaisen avustavan käytön lisäksi tekoäly voi toimia itsenäisesti ja tehdä käytännössä päätöksiä. Se, millaisia päätöksiä se tekee, perustuu kuitenkin opittuihin assosiaatioihin niistä ihmisen toimintaan perustuvista malleista, joita algoritmin koulutuksessa on käytetty. Tällainen teknologia voi automatisoida monia palveluammatteihin ja toimistotyöhön liittyviä rutiinitoimia ja käytännössä korvata niin sanottuja valkokaulusammatteja. Lisäksi tekoälyä käytetään ”digitaalisena filtterinä”, joka tekee päätöksiä siitä, mitä käyttäjät näkevät hakukoneissa, sosiaalisessa mediassa tai suoratoistopalvelujen suosituksissa.

Viime aikoina huomiota on herättänyt erityisesti generatiivinen tekoäly, joka jäljittelee ihmisen luomisprosessia: erilaiset tekstiä tuottavat kielimallit, kuvia sanallisen kuvauksen pohjalta tekevät mallit ja viimeisimpänä lyhyitä realistisia animaatioita tekevät algoritmit. Näistäkin on huomattava, että ne lähinnä toistavat ja varioivat niitä rakenteita, joita ne ovat oppineet siitä opetusmateriaalista, jota niiden luomiseen on käytetty. Tämä riittää kuitenkin melko moneen sellaiseen asiaan, johon on aiemmin tarvittu ihmistä.

Eettisiä riskejä

Tekoälyteknologiaan liittyy monia eettisiä riskejä. Osa näistä liittyy itse teknologiaan, osa siihen, miten sitä käytetään. Esimerkiksi yksityisyyteen ja valvontaan liittyvät kysymykset ovat herättäneet huolta. Digitalisaatio ja ihmisten toiminnallaan tuottama data itsestään (esimerkiksi erilaisissa tietojärjestelmissä, sosiaalisessa mediassa, nettikaupoissa ja hakukoneissa) ovat mahdollistaneet massiivisen datan keruun ja varastoinnin, mutta tekoäly on mahdollistanut myös sen tehokkaan analyysin ja säännönmukaisuuksien löytämisen. Tätä hyödynnetään esimerkiksi tuotekehittelyssä keräämällä käyttäjädataa, mutta tekoäly mahdollistaa myös tehokkaan automatisoidun tarkkailun. Eettinen peruskysymys liittyy yksilön oikeuteen määrätä, kuka hänestä tietää ja mitä. Sekä kaupalliset että julkiset toimijat yleensä kysyvät käyttäjiltä suostumuksen, mutta on kyseenalaista, missä määrin ihmiset kykenevät antamaan informoidun suostumuksen omasta datastaan, kun datan kerääjäkään ei välttämättä vielä tiedä, mitä kaikkea sillä voidaan tulevaisuudessa tehdä. Tekoäly mahdollistaa myös manipulaation. Data-analyysi mahdollistaa tehokkaan täsmämarkkinoinnin ja kohdistetun disinformaation levittämisen. Kuvien, äänen ja videokuvan syväväärennökset ovat tulleet jäädäkseen luomaan sekaannuksia.

Päätöksiä tekevän tekoälyn kohdalla huolenaiheeksi nousevat läpinäkyvyys ja luotettavuus. Emme aina tiedä, miten algoritmi päätyy tulokseensa, eikä se anna perusteluja. Joskus assosiaatiot voivat perustua täysin irrelevantteihin yhteyksiin. Klassinen esimerkki on algoritmi, joka oppi erottamaan suden ja koiran toisistaan kuvista – koska kaikissa opetuskuvissa ja testikuvissa susikuvissa oli lunta mutta koirakuvissa ei, ja assosiaatio perustui tähän. Yksinkertaiset virheassosiaatiot paljastuvat helpommin, mutta monimutkaisemmissa tilanteissa tiedollinen luotettavuus kärsii. Jos algoritmin toiminta on läpinäkymätöntä, emme voi varmistaa ja korjata prosesseja. Herää myös kysymys, kuka on vastuussa päätöksenteossa esiintyvistä virheistä tai epätoivotusta toiminnasta.

Yksi lähde virheille ja epätoivotulle tuloksille on opetusdata: sen vinoumat (kuten vaikkapa rasistiset ennakkoluulot) periytyvät algoritmin toimintaan ja voivat jopa vahvistua. Opetusdatan valinnasta vastaavia ihmisiä voi ehkä tässä tilanteessa pitää eettisessä vastuussa. Myös virheet algoritmien suunnittelussa tai niiden väärinkäyttö tuovat moraalisen vastuun ihmistoimijoille. Tilanne on kuitenkin monimutkaisempi, jos puhumme autonomisista järjestelmistä kuten itseohjautuvat autot, autonomiset asejärjestelmät tai vaikkapa tukihakemuksia käsittelevät järjestelmät. Ne voivat toimia hyvin konteksteissa, joihin ne on suunniteltu, mutta arvaamattomasti toisissa. Käytön aikana oppivat järjestelmät voivat myös päätyä sellaiseen epäeettiseen toimintaan, josta on vaikea pitää vastuussa ketään ihmistä.

Voiko tekoälystä tehdä eettisen?

Itsenäisesti toimivien järjestelmien tulevaisuuden näkymät ovat synnyttäneet uuden soveltavan etiikan osa-alueen, kone-etiikan. Sen tavoitteena on teknologian suunnittelu niin, että sovellukset toimivat käyttöyhteydessään sellaisilla tavoilla, joita pidämme eettisesti oikeina. Tavoitteena on ratkaista, mitkä ovat eettiset tavat toimia missäkin tilanteessa. Jos tekoäly tekee itsenäistä päätöksentekoa vaihtuvassa joukossa tilanteita, sääntöjen on oltava yleisempiä kuin täysin tilannekohtaisia: emme voi ennakoida jokaista tilannetta etukäteen. Tähän ongelmaan törmätään esimerkiksi jo itseohjautuvissa autoissa. Toisin sanoen autonomiselle tekoälylle täytyisi opettaa moraalisääntöjä. Mutta miten tämä onnistuu?

Tekoälylle on yritetty opettaa eettisiä sääntöjä ”top-down”-mallisesti (säännöt ensin ja niille tulkinta) ja ”bottom-up”-mallisesti (sääntöjen koneoppiminen valikoimalla oikeat valinnat), mutta molemmat ovat osoittautuneet toimimattomiksi. Myös erilaisia näitä menetelmiä yhdistäviä hybridimalleja on kehitetty, mutta nekään eivät ole tuottaneet tulosta. Jos moraaliin ei liity mitään yliluonnollista, on periaatteessa täysin mahdollista replikoida tai ainakin simuloida siihen liittyvää ajattelua myös keinotekoisesti. Voi kuitenkin olla, että emme kykene tällä hetkellä luomaan järjestelmiä, joilla olisi moraalikognition jäljittelyyn vaadittavia valmiuksia.

Tästä palaamme jälleen moraalisen vastuun kysymykseen. Kuka on vastuussa virheistä ja odottamattomasta toiminnasta? Algoritmi ei ole moraalinen toimija. Algoritmin suunnittelija ei voi olla vastuussa sellaisesta algoritmin käyttäytymisestä, joka on ennakoimatonta. Algoritmin käyttäjä puolestaan ei tunne algoritmin toimintaa niin hyvin, että voisi olla vastuussa sen käytön kaikista seurauksista. Mutta jos kukaan ei ole eettisessä vastuussa, vastuu siirtyy siihen, että teknologia on ylipäätään otettu käyttöön tavalla, jolla se on otettu käyttöön: vastuukysymyksistä kannattaa siirtyä vastuullisen suunnittelun kysymyksiin. Haluamme ehkä maksimoida teknologian turvallisuuden ja luotettavuuden käyttökonteksteissaan jo suunnittelussa. Lisäksi riskit on hyvä huomioida kontekstuaalisesti: käyttökontekstit tulee valita niihin sisältyvien riskien mukaan.

Arvosensitiivinen suunnittelu

Niin sanotussa ”arvosensitiivisessä suunnittelussa” huomioidaan teknologian käyttökontekstin arvot suunnittelussa, kehittelyssä ja käyttöönotossa. Yleensä relevantteihin arvoihin sisällytetään niin moraaliset kuin kulttuurisetkin arvot, säännöt ja periaatteet. Pyrkimyksenä on teknologian sujuvampi integraatio ja toiminnallisten haittojen minimointi käyttökonteksteissaan, mutta siihen sisältyy myös eettisten riskien minimointi ja käytäntöjen eettisyyden pohtiminen. Arvosensitiivisessä suunnittelussa selvitetään sekä sen käytäntöjen kokonaisuuden toiminta, johon teknologia tuodaan, että näissä käytännöissä vaikuttavat arvot. Esimerkiksi hoivarobotteja suunnitellessa selvitetään ne hoitokäytännöt, joihin robotti tuodaan: mikä on yksittäisen käytännön design ja suhde muihin käytäntöihin, automatisoitavan elementin (esimerkiksi potilaan nostaminen sängyltä pyörätuoliin) funktio kokonaisuudessa, ja mitä “toissijaisia” funktioita käytännöllä on. Esimerkiksi potilaan ja hoitajan jutustelu on toissijainen suhteessa suoritettavaan toimeen, mutta sillä voi olla silti merkitystä sekä potilaalle (ihmiskontakti) että hoitotyön onnistumiselle (potilaan vointia koskevan informaation välittyminen). Analysoitavia arvoja ovat ne arvot, jotka ohjaavat hoitokäytäntöjä: esimerkiksi potilaan ja hoitohenkilökunnan tarpeet, hoitajan kyky huomata potilaan tarpeet ja vastata niihin oikealla tavalla, potilaan itsemääräämisoikeus, itsenäisen toimintakyvyn maksimointi, yksityisyys ja ihmisarvon tunne. Hoivarobotin eettinen käyttöönotto edellyttää hoitokontekstin kokonaisuuden arviointia (esimerkiksi organisaatio ja resurssien jakaantuminen) ja sen eri osapuolten huomioimista (potilaat, hoitajat, lääkärit, omaiset, hallinto), toiminnan kokonaisuuden ja sitä ohjaavien arvojen ymmärtämistä ja näissä tapahtuvan muutoksen ymmärtämistä, jos teknologia otetaan käyttöön. Erityisesti on tärkeää, että kaikki osapuolet tulevat huomioitua.

Tekoälyteknologia liittyy kuitenkin laajempaan yhteiskunnalliseen muutokseen ihmisten elämäntavoissa, sosiaalisissa suhteissa ja käytännöissä, työelämässä, instituutioiden toiminnassa, talouden rakenteissa ja niin edelleen. Kyse ei ole myöskään vain olemassa olevien käytäntöjen hallitusta muutoksesta, vaan uusien käytäntöjen ja rakenteiden syntymisestä. Tämä tarkoittaa, että arvosensitiivistä suunnittelua tarvitaan myös laajemmalla yhteiskunnallisella tasolla: muutoksesta, yhteiskunnallisista tavoitteista ja eettisten riskien minimoimisesta olisi järkevää käydä kattavaa arvokeskustelua nyt, kun teknologisen kehityksen suunnat ovat vasta muotoutumassa.

Tomi Kokkonen
Filosofi, FT, VTM
Robophilosophy, AI Ethics and Datafication Research Group
Helsingin yliopisto