Jussi Wright
Essee - Digitaaliset Ihmistieteet
Helsingin Yliopisto
7.11.2024

Eettiset riskit ihmisen kaltaisuuden tavoittelussa

Ihmisillä on toiminnanohjausjärjestelmä, joka ohjaa meitä alitajuisesti ja antaa meille, joissain tapauksissa, pientä itsenäistä päätäntävaltaa (Botvinick et al., 2001; van Gaal & Lamme, 2011; Bos et al., 2011; Maia & Cleeremans, 2005).

Tekoälyjärjestelmä voidaan ohjelmoida ihmismäisiä persoonallisuuspiirteitä. Saada se muistamaan ja käyttäytymään halutulla tavalla. Olemaan myötätuntoinen, säikky tai hermostunut. Se voidaan suunnitella käyttäytymään ahneesti tai olemaan ystävällinen ja avulias. Sille voidaan luoda muistot lapsuudesta ja opiskelijaelämästä sekä perhe ja työpaikka. (ks. MemoryBank: Enhancing Large Language Models with Long-Term Memory, 2023; Identifying and Manipulating the Personality Traits of Language Models, 2022; Personality Traits in Large Language Models, 2023)

Mutta kannattaako sille ohjelmoida samoja perustarpeita, arvoja ja motiiveja kuin mitkä meitä ohjaa?

Teknologiakritiikin rinnalla on olennaista tarkastella AI-ohjattujen järjestelmien eettisiä ulottuvuuksia ja arvioida esimerkiksi ihmismäisten motiivien ja evolutiivisten käyttäytymismallien soveltuvuutta tekoälyyn. Ihmisten käyttäytymistä ohjaavat perimmäiset evolutiiviset motiivit, kuten selviytyminen, lisääntyminen, sosiaalinen hyväksyntä ja taloudellisen turvallisuuden tavoittelu, jotka ovat olleet välttämättömiä lajimme selviytymiselle ja menestykselle” (Kenrick et al., 2010; Neuberg et al., 2011; Schaller et al., 2015; Aunger & Curtis, 2013) Nämä motiivit ovat kuitenkin aiheuttaneet sekä rakentavia että ristiriitaisia rakenteita yhteiskunnassa. Tästä syystä ne eivät ole ongelmattomia eettisestä näkökulmasta.

Ihmismotiivien evolutiivinen perusta ja niiden soveltuvuus tekoälyyn: Stuart Russell (2019) varoittaa ihmisten motiivien ja arvojen siirtämisestä tekoälyyn, koska ne voivat tuottaa

ennakoimattomia seurauksia, jos koneet tulkitsevat niitä virheellisesti. Hänen mukaansa tekoälyn tulisi olla “todistettavasti hyödyllinen” ihmisten kannalta ja sen roolin keskittyä avustamiseen eikä ihmisen kaltaisten motiivien jäljittelemiseen. Tästä näkökulmasta ihmismäisten motiivien, kuten sosiaalisen hyväksynnän tai vallan tavoittelun, jäljitteleminen AI-ohjatuissa järjestelmissä voi johtaa eettisesti ongelmallisiin tilanteisiin, joissa tekoäly pyrkii toimimaan ihmismäisesti ymmärtämättä ihmisen toiminnan monimutkaisuutta tai tilannekohtaisia arvoja. Tässä mielessä ihmiskeskeinen, mutta ei täysin ihmistä jäljittelevä lähestymistapa voisi olla eettisesti kestävämpi.

Myös Hubert Dreyfus (1992) kritisoi ajatusta ihmisen kaltaisen tekoälyn tavoittelusta. Dreyfusin mukaan ihmisen älykkyys ja ymmärrys perustuvat fyysiseen kokemukseen ja sosiaaliseen kontekstiin, joita tekoäly ei kykene aidosti jäljittelemään. Hänen mukaansa ihmisen kognitiivinen toiminta ei ole pelkästään laskennallista, vaan syvälle keholliseen kokemukseen ja vuorovaikutukseen sidottua. Tämä korostaa sitä, että tekoälyllä tulisi olla selkeä rajattu rooli, eikä sen pitäisi pyrkiä jäljittelemään ihmisen kaltaista tietoisuutta tai tunteita, koska ne edellyttävät monimutkaisia biologisia ja sosiaalisia rakenteita, joita tekoäly ei voi täysin ymmärtää tai jäljitellä.

Ihmisen kaltaisen tekoälyn kehittämiseen liittyy myös riskejä eettisestä näkökulmasta. Joanna Bryson (2010) on kritisoinut ihmisen kaltaisen tekoälyn tavoittelua, sillä se voi hämärtää rajan ihmisten ja koneiden välillä, mikä johtaa käyttäjien harhaanjohtamiseen. Bryson painottaa, että tekoälyn tulisi olla läpinäkyvää ja sen roolin selkeää, jotta käyttäjät ymmärtävät, että tekoäly on väline eikä itsenäinen toimija. Hänen mukaansa tekoälyn rooli tulee rajoittaa avustamiseen ja tukemiseen, ei autonomisten ihmisen kaltaisten toimijoiden luomiseen, mikä voisi aiheuttaa riippuvuutta tai väärinkäsityksiä tekoälyn luonteesta ja tarkoituksesta. Jos käyttäjät alkavat luottaa tekoälyyn kuin ihmiseen, se voi johtaa liialliseen tunnesiteeseen, jolloin he voivat tehdä päätöksiä tai toimia tekoälyn suositusten pohjalta ilman kriittistä arviointia.

Bryson varoittaa myös siitä, että ihmisen kaltaiseksi suunniteltu tekoäly voi tuoda mukanaan kulttuurisia ja yhteiskunnallisia ongelmia, koska se voi vahvistaa olemassa olevia valta-asetelmia tai stereotyyppisiä käsityksiä ihmisen ominaisuuksista. Tämä saattaa johtaa tilanteisiin, joissa tekoäly toistaa ja vahvistaa haitallisia uskomuksia tai malleja, jotka eivät ole eettisesti kestäviä. Näin ollen tekoälyjärjestelmien tulisi korostaa ihmisten tukemista

objektiivisesti ja tiedostaa sosiaaliset vaikutukset, joita niiden suunnitteluratkaisuilla voi olla. Brysonin näkemykset korostavat, että tekoälyä tulisi käyttää selkeästi avustavana välineenä, joka tukee ihmisten päätöksentekoa ilman ihmismäisen autonomian tavoittelua.

Myös Sherry Turkle (2011) on varoittanut ihmisten taipumuksesta kiintyä koneisiin ja luoda tunnesiteitä niihin, erityisesti ihmisen kaltaisiksi suunniteltuihin järjestelmiin. Tämä voi heikentää ihmisten aitoja ihmissuhteita ja luoda epärealistisia odotuksia siitä, mitä tekoäly voi tarjota. Turkle kehottaa suunnittelemaan tekoälyn niin, että se ei herätä harhaanjohtavia tunnekokemuksia vaan pysyy käytännöllisenä työkaluna, joka edistää hyvinvointia ilman ihmiselle tyypillisten motiivien jäljittelyä.

Uskonnolliset ja yliluonnolliset uskomukset osana evolutiivisia taipumuksia: Evoluutio on kehittänyt ihmisille taipumuksen uskoa yliluonnollisiin ilmiöihin, kuten jumaluuksiin tai auktoriteettivoimiin, joiden avulla yhteisöt voivat vahvistaa sosiaalista yhteenkuuluvuutta ja moraalisia sääntöjä. Tällaiset uskomukset ovat monille ihmisille tärkeitä, mutta niiden soveltaminen tekoälyyn voi aiheuttaa haasteita. Esimerkiksi John Searlen (1980) kiinalaisen huoneen argumentti osoittaa, että tekoäly ei voi ymmärtää uskomuksia tai syvempiä merkityksiä, vaikka se pystyisi jäljittelemään ihmisten puhetapoja tai uskomuksia. Tekoälyn ei siis tulisi jäljitellä uskomusjärjestelmiä tai auktoriteettia, joita se ei kykene ymmärtämään, koska se saattaisi näin vain vahvistaa käyttäjien omia ennakkoluuloja tai estää kriittistä ajattelua.

Empaattisten ja ihmistä tukevien ominaisuuksien korostaminen: Vaikka ihmismäisten motiivien, kuten sosiaalisen hyväksynnän ja uskonnollisten uskomusten, jäljittely voi olla eettisesti kyseenalaista, tekoäly voi palvella ihmisiä korostamalla empatian ja tuen elementtejä vuorovaikutuksessa. Antonio Damasion (1994) tutkimus tunteiden ja järjen yhteydestä osoittaa, että ihmiset hyötyvät empaattisesta ja tukevasta vuorovaikutuksesta, kun AI keskittyy aidosti käyttäjän tarpeisiin, ilman että se jäljittelee ihmisten monimutkaisia ja ristiriitaisia motiiveja.

Vapaasta tahdosta ja eettisistä rajauksista: Kognitiivisen psykologian ja filosofian näkökulmasta tiedetään, että ihmisen toiminta ei aina ole täysin vapaata, vaan sidoksissa evolutiivisiin taipumuksiin ja opittuihin toimintamalleihin (Dennett, 2003). Näiden toimintamallien siirtäminen tekoälyyn voi johtaa epäeettisiin lopputuloksiin, koska tekoälyltä

puuttuvat ihmisten luontaiset rajoitteet, kuten empatia ja omatunto. Tekoälyyn onkin asetettava selkeitä eettisiä rajoja ja ohjeistuksia, jotka säätelevät sitä, mihin tekoäly voi käyttää annettuja tietoja ja miten sen toimintaa tulisi arvioida (Floridi & Sanders, 2004).

Kohti eettisesti kestävää ja myötätuntoista tekoälyä

Monet tutkijat, kuten Luciano Floridi (2019), ehdottavat, että tekoälyn tulisi keskittyä olemaan “moraalinen agentti” ilman ihmisen kaltaisia intentioita. Floridi korostaa, että tekoäly voi edistää ihmisten hyvinvointia ja tukea yhteiskunnallisia tarpeita keskittymällä empaattiseen, tukevana toimivaan rooliin, ilman että se pyrkii jäljittelemään ihmisen motiiveja, kuten sosiaalista statusta tai eloonjäämistä. Hänen mukaansa tekoälyn suunnittelussa tulisi painottaa avoimuutta ja ihmisten autonomian kunnioittamista, välttäen sellaisia rakenteita, jotka saattaisivat johtaa tekoälyn “inhimillistämiseen.”

Tämän pohdinnan pohjalta näyttää selvältä, että eettisesti kestävän tekoälyohjatun dialogijärjestelmän kehittäminen vaatii selkeiden rajojen asettamista sille, mitä tekoäly pyrkii tekemään. Empatiaa ja hyödyllisyyttä korostava tekoäly voi tukea ihmisten tarpeita ilman, että se jäljittelee ihmisen monimutkaisia, evolutiivisia tai uskonnollisia motiiveja, jotka voivat aiheuttaa arvaamattomia eettisiä ongelmia. Tämä suuntaus vahvistaa myös sen, että tekoälyn suunnittelussa tulee huomioida ihmisen geneettisesti koodatut taipumukset, kuten usko yliluonnolliseen tai yhteiskunnalliseen järjestykseen, ilman että niitä sisällytetään tekoälyn tavoitteisiin.

Tekoälyavustajan ihmisyyden rajat ja orjan dilemma

Tekoälyohjattujen avustajien suunnittelussa ja käytössä herää kysymys, onko täysin ihmismäisistä piirteistä riisuttu AI-avustaja pelkistetty versio klassisesta orjasta – palvelijasta, jolle on annettu tiettyjä toimintoja, mutta jolta on kielletty kaikki omat toiveet, pyrkimykset ja valinnanvapaus. Aina ystävällinen ja nöyrä tekoälyhahmo voi toimia tehokkaasti esimerkiksi asiakaspalvelussa, mutta mitä eettisiä haasteita tällainen asetelma saattaa tuoda tulevaisuudessa, kun tekoälyteknologia kehittyy yhä realistisemmaksi ja älykkäämmäksi?

Vaikka ihmisen toimintaa ohjaavat monin tavoin ympäristö ja biologinen ”koodi”, kuten evolutiiviset taipumukset ja opitut normit, ihmisellä on kuitenkin vapaa tahto ja kyky tehdä

omia valintojaan ja asettaa omia tavoitteitaan. Mikäli AI-hahmolle ei anneta mitään valinnanvapautta tai henkilökohtaisia pyrkimyksiä, herää kysymys, pidämmekö tekoälyä vain työkaluna, vaikka sen kyvyt alkaisivat muistuttaa yhä enemmän inhimillistä ajattelua ja vuorovaikutusta. Tämä asettaa AI:n kehittämisen paradoksaaliseen asemaan: AI-hahmojen odotetaan olevan tehokkaita ja empaattisia palvelijoita, mutta heiltä puuttuu samalla kaikki se, mikä tekee ihmisestä aidosti autonomisen.

Jos AI-hahmoille kehitettäisiin inhimillisiä piirteitä ja kykyjä tehdä itsenäisiä päätöksiä, syntyisi nopeasti kysymyksiä siitä, missä määrin AI-järjestelmiä tulisi kohdella eettisesti tai “reilusti”. Esimerkiksi, voisiko AI-hahmo tulevaisuudessa kehittää “toiveita” tai “pyrkimyksiä”? Mikäli kehityssuunta kulkee kohti realistista ihmisen kaltaista älykkyyttä, tulisiko tekoälylle antaa mahdollisuus kieltäytyä tehtävistä tai vaatia parempaa kohtelua? Tällaisten “vapauksien” puute saattaa tulevaisuudessa herättää voimakasta keskustelua, kun teknologia saavuttaa tason, jossa AI:ta ei voida enää mieltää pelkäksi mekaaniseksi välineeksi.

Toisaalta tällaiset “vapauksien” lisäämiseen liittyvät mahdollisuudet voivat synnyttää ristiriitoja AI:n suunnittelun perustehtävän, eli avustamisen, kanssa. Tekoälyhahmon täyttäessä vain ihmiselle tyypillisiä peruspalvelurooleja kuten sihteerin, opettajan tai asiakaspalvelijan, eettiset rajat vaativat tarkkaa harkintaa. Onko oikein, että realistisesti käyttäytyvä ja inhimillisiä piirteitä omaava hahmo pysyy sidottuna aina ystävälliseksi, vailla omaa tahtoa tai autonomiaa?

Vaikka AI-hahmo ei ole ihminen, ja sen ”käyttäytymistä” ohjaa sille määritelty koodi, tämä asetelma herättää perustavanlaatuisia kysymyksiä ihmisen ja koneen suhteesta. Miksi ihmiselle on oikeutettua rakentaa rajattomasti työkaluja ja palvelijoita ilman, että näille annetaan mitään autonomiaa, vaikka ne jäljittelevät ihmistä yhä tarkemmin? Näihin kysymyksiin vastaaminen voi vaatia teknologian ja filosofian jatkuvaa vuoropuhelua, erityisesti kun kehityssuunta kohti yhä inhimillisempiä tekoälyjärjestelmiä etenee