Tulevaisuusvaliokunnan kokeilu tekstiä tuottavasta ja dialogisesta tekoälystä tukiälyn näkökulmasta

Risto Linturi, tulevaisuudentutkija

Tulevaisuusvaliokunnan kokeilu tekstiä tuottavasta ja dialogisesta tekoälystä tukiälyn näkökulmasta

Mistä on kyse?

Tekoäly on tutkimusalueena hyvin laaja, ja näkökulmat ulottuvat ihmiselämän, yhteiskunnan ja talouden kaikkiin osa-alueisiin. Tässä artikkelissa keskitytään yleisen tukiälyn näkökulmaan eli siihen, miten yksilö voisi käyttää tekoälyä avukseen yleisissä tai yllätyksellisissä tilanteissa. Pääpaino tarkastelussa kohdistuu luonnollista kieltä käyttävien dialogisten järjestelmien käyttöön. Joitakin huomioita kohdistetaan myös konenäön ja emotionaalisen tekoälyn tukiäly-tyyppisiin sovelluksiin. Artikkeli taustoittaa tulevaisuusvaliokunnan suorittaman kokeilun, jossa demonstroitiin ja testattiin merkittävän tekstiä generoivan tekoälyn kykyä dialogisessa käytössä sekä käyttäjän mahdollisuutta vaikuttaa järjestelmän toimintatapaan ja sen eettisiin valintoihin.

Ihmiset käyttävät runsaasti apuvälineitä muistin tukemiseen, tiedon käsittelyyn ja -hakuun sekä havaintojen tekemiseen. Tukiäly-käsite ei varsinaisesti tarkoita tätä nykyään arkista tietotekniikkaa vaan tekoälyn soveltamista merkitysten ja mahdollisuuksien hahmottamiseen. Käsite on vakiintumaton, mutta sisältönä tukiälyyn viitattaessa on tavallisesti keskustelunomainen apu joko asiantuntijan omassa työssä tai ihmisen arjessa. Tukiälyksi voidaan katsoa myös aisteiltaan ja muistiltaan rajoittuneiden ihmisten käyttöön soveltuva keskusteleva ja ympäristöä havainnoiva tekoäly.

Mihin ihminen sitten keskusteluapua tavallisimmin tarvitsee? Tietokoneilta on totuttu odottamaan täsmällisiä ja oikeita vastauksia tai virheetöntä tiedon järjestelyä. Toiselta ihmiseltä ei tällaista virheettömyyttä vaadita ja siltikin keskustelut nähdään hyödyllisiksi. Keskusteluilta odotetaan näkemysten avartamista, asioiden erilaisten merkitysten havaitsemista ja uhkien sekä mahdollisuuksien näkemistä. Kyse ei siis useinkaan ole siitä, että keskustelukumppanin tulisi osata antaa faktuaalinen tai virheetön vastaus selkeään kysymykseen vaan siitä, että keskustelu edistää ymmärrystä monitahoisen asian merkityksiin, joita kukin sitten voi selvittää eri keinoin tarkemmin. Tällaisen keskusteluavun hyöty on vaikea arvioida tilastollisin keinoin, mutta hermeneuttiset tutkimusmenetelmät soveltuvat hyödyn arviointiin erityisen hyvin.

Ihmiset ovat keskustelukumppaneina hyvin vaihtelevia. Tilanteen lisäksi valitsemamme rooli, tietomme ja tunteemme sekä kokemuksemme vaikuttavat tapaamme reagoida asioihin. Yleinen käsitys tekoälystä on ollut, että se olisi tässä suhteessa yksinkertaisempi ja sen päätelmiin vaikuttaisi suoraviivaisemmin annettu opetusaineisto, että esimerkiksi epätasapainoinen opetusaineisto johtaisi tekoälyn vääristyneisiin ennakkoasenteisiin. Toisaalta emme odota, että havaintojen tilastollinen vääristymä tekisi ihmisistä vääjäämättä esimerkiksi rasisteja.

Näiden kysymysten tutkiminen on tullut mielekkääksi tekstiä generoivan tekoälyn kehityttyä riittävästi, että keskustelu vaikuttaa sujuvasti liikkuvan inhimillisen elämän erilaisilla merkitystasoilla. OpenAI-yhtiön kehittämä GPT-3 on ensimmäinen järjes-

telmä, jonka on yleisesti katsottu täyttävän tämän ehdon. Kyse on niin kutsutusta kielimallista, joka ennustaa, mikä sana parhaiten soveltuu aiemman tekstin jatkoksi. Arvioinnissa otetaan kieliopin lisäksi huomioon viitekehys ja hahmotetaan vastaavuuksia opetusaineistoon. Tämä johtaa käytännössä siihen, että GPT-3 näyttää huomioivan myös emotionaaliset konnotaatiot.

Parhaiten GPT-3 tulee ymmärretyksi, kun sitä ajatellaan näytelmäkirjailijana. Mikäli pyydetään jatkoa tekstiin, jonka mukaan kuu on juustoa, tulkitsee GPT-3 viitekehyksen saduksi ja saattaa jatkaa kuvaamalla lentäviä lehmiä. Vastauksien todenperäisyyteen ei siis voi luottaa sen paremmin kuin ihmistenkään antamien vastausten todenperäisyyteen. Tässä on huomattava, että GPT-3:n opetusaineisto on erittäin laaja. Wikipedia on noin 3% kaikesta sen opetukseen käytetystä materiaalista. Joukossa on laajasti kertomakirjallisuutta, tieteellisiä artikkeleita ja verkkokeskusteluita. Kuten aiemmasta voi havaita, eivät GPT-3:n kirjoittamat ajatukset ensisijaisesti määräydy opetusaineiston tilastollisten painotusten mukaan vaan annetun syötteen vaikutus on ensisijainen.

Tärkeäksi tutkimuskohteeksi osoittautuu se, millä tavoin GPT-3:n kaltaista kielimallia voidaan ohjata toimimaan halutulla tavalla. Voidaanko syötteeksi antaa kuvaus tilanteesta ja sen lisäksi ihmisen kaltainen persoonallisuus ja rooli, jonka mukaisessa viitekehyksessä vastauksia halutaan. Tähän kysymykseen liittyvän 60 dokumentoidun kokeilun sarjaan merkittävänä osana liittyi myös jäljempänä tässä raportissa yksityiskohtaisesti raportoitu tulevaisuusvaliokunnan suorittama kokeilu.

Tulevaisuusvaliokunnan kokeilussa käytettiin Project December-alustaa, joka muuttaa GPT-3:n toimimaan dialogisessa muodossa ja tarjoaa helpon mahdollisuuden roolihenkilön ja tilanteen kuvaamiseen, joiden puitteissa vastaukset halutaan. Kokeilussa tilanne kuvattiin eduskunnan asiantuntijan kuulemiseksi. Persoonan vaikutus vastauksiin havainnollistettiin esittämällä samat kysymykset kahdelle eri tavoin määritellylle persoonalle. Toisen, Muskien, roolimallina oli Elon Musk ja toisen, Saaran, Greta Thunberg, opiskelunsa päättäneenä ympäristöpolitiikan asiantuntijana.

Kokeilu osoitti GPT-3:n tuottavan valittujen roolien mukaisia vastauksia erittäin johdonmukaisesti. Repliikit olivat pääsääntöisesti hyviä vastauksia annettuihin kysymyksiin, mutta roolikohtaisesti hyvin erilaisia. Samoihin kysymyksiin Muskie ehdotti teknologisesti kunnianhimoisia, visionäärisiä hankkeita ja arvioi kehitystä teollisuusjohtajan näkökulmasta. Saara korosti ympäristöpolitiikkaa ja jäsensi puheitaan ympäristöpoliittisen kielen mukaan. Erot sopivat persoonallisuuksiin jopa huumoria testattaessa. Roolien mukaiset arvot ja osaamispainotukset näkyivät siis yhtä selvästi kuin ne tyypillisten asiantuntijoina kuultujen eri taustaisten ihmisten vastauksissa näkyvät. Tästä on helppo päätellä, että GPT-3 on erittäin joustava eikä sillä itsessään ole varsinaista jäykkää asennetta. Näytelmäkirjailijan kaltaisesti vastaus sovitetaan annettuun tilannekuvaukseen, roolikuvaukseen ja aiempaan dialogiin niissä puitteissa, joissa opetusaineisto antaa vastauksiin riittävästi virikkeitä.

Kokeilussa kävi ilmeiseksi, että GPT-3:n laaja opetusaineisto on kyselijän verrattain helposti ja selkeästi saavutettavissa vastaavalla tavalla kuin laajasti lukeneen ihmisen tietomäärä. Kysymyksiin saatiin hyviä, kohdistettuja vastauksia, jotka jatkokysymysten avulla tarkentuivat ja, joista saattoi keskustella. Tekoälyn kanta saattoi myös muuttua

jatkokysymysten ja keskustelussa esitettyjen väitteiden myötä. Ilmeistä oli sekin, että vastaukset saattoivat osittain olla keksittyjä ja osittain tiedollisesti epämääräisiä.

Selvältä kokeilujen jälkeen vaikutti, että GPT-3 dialogisessa käytössä tuottaa merkittävästi erilaisia hyötyjä verrattuna nykyisiin internetin hakukoneisiin. Selväksi myös kävi, ettei kyse ole luotettavana pidettävästä tietolähteestä vaan virikkeellisestä ja näkökulmia avartavasta keskustelijasta, kuten oletus olikin.

Tekoäly muutosvoimana, kaksi polkua tulevaisuuteen:

Tulevaisuusvaliokunnan GPT-3 -kokeilu on tarkemmin taustoitettu ja dokumentoitu tämän jälkeisessä luvussa. Kokeilun jälkeen kehitys on edennyt useilla merkittävänä pidettävillä tavoilla.

Kiinalaistutkijat ovat kertoneet GPT-3:a laajemmasta kielimallista, joka kykenee tekstin tuottamisen lisäksi tulkitsemaan kuvia. Myös NVidia ja Microsoft ovat julkaisseet yhteistyössä kehittämänsä GPT-3:a laajemman kielimallin, joka heidän omissa testeissään tuottaa tätä parempia tuloksia. Siinä käytetyt algoritmit ovat avoimesti saatavilla. Avoimen koodin yhteiskehittämisen tuloksena on lisäksi tullut saataville melko laajoja kielimalleja opetusaineistoineen. Uutena kehityssuuntana on malli, joka kykenee muistamaan tiivistelmiä eri henkilöiden kanssa käymistään keskusteluista.

Huomion arvoista on myös, että OpenAI on nyttemmin estänyt tässä kuvatun kokeilun kaltaisen käytön vedoten tekoälyn riskeihin. Yhtiö ilmoitti, ettei se enää salli pääsyä GPT-3 -palveluun, joka olisi yleisölle avoin, vapaasti roolitettavissa ja sensuroimaton.

Tulevaisuusvaliokunnan kokeilu ja sen jälkeinen OpenAI:n reaktio sekä hieman saman suuntainen EU:n tekoälyn sääntelyyn liittyvä ehdotus, jota kommentoin myöhemmin, osoittavat kaksi hyvin erilaista polkua tulevaisuuteen. Kutsuttakoon toista **isoveljen ohjaamaksi tekoälyksi** ja toista **pikkusisarusten tukiälyksi**.

Isoveljen ohjaamalla tekoälyllä tarkoitetaan tässä maailmaa, jossa merkittävin investoinnein on synnytetty keskustelevia tekoälyjä siten, että niiden persoonat, tavoitteet ja arvomaailmat on keskitetysti suunniteltu ja reguloitu toisaalta palvelemaan kehittäjiensä tarkoitusperiä ja toisaalta noudattamaan yhteiskunnan asettamia reunaehtoja. Tämän suunnan etuna on se, että tekoälystä voidaan yrittää tehdä turvallinen. Turvallisuudella tarkoitetaan, ettei se esimerkiksi kävisi kenenkään kanssa keskustelua, joka olisi omiaan asettamaan tämän vaaraan tai yllyttäisi rikoksiin.

Ongelmalliseksi isovelimallissa voidaan nähdä yksimielisyyden puute siitä, keiden arvot ja tavoitteet nähdään suotaviksi syöttää tekoälylle. Toisena ongelmana voi hyvästä syystä pitää sitä, että tällainen ”hyväntahtoinen tekoäly” voi osoittautua mitään oikeasti ymmärtämättömäksi tolloksi, jonka kanssa on vaikea keskustella ongelmallisista asioista. Esimerkkinä toimii Facebookin käyttämä, ilmiantoja tarkistava sensuuri, jonka harhauttaminen on ilmeisen helppoa sekä hyvässä että pahassa. Toistaiseksi ei tällaista virheettömästi ja yleispätevästi hyväntahtoista, mielekkääseen laajaan ja vapaaseen keskusteluun kykenevää tekoälyä ole onnistuttu kehittämään.

Pikkusisarusten tukiäly -mallissa tekoälypalveluita kehitetään kokeilun osoittamalla tavalla geneerisiksi. Käyttäjän itsensä on mahdollista kuvailla, minkälaisen persoonan neuvoja tai keskusteluapua kulloinkin kaipaa. Kokeilut osoittavat selkeästi, miten helposti jo nykyisen tasoisen GPT-3:n ja verrattain vähäisen kokemuksen avulla tämä määrittely on tehtävissä onnistuneesti. Etuna lähestymistavassa on, että tukiäly tulkitsee opetusaineistoaan käyttäjänsä määrittämän viitekehyksen ja arvomaailman mukaan. Tämän kaltainen luottamus käyttäjään on ollut vapaan yhteiskunnan perinne.

Tukiälyn voi tässä skenariossa asettaa keskustelemaan vaikkapa huumeriippuvan tai peloista kärsivän ihmisen roolissa ja opetella ymmärtämään mahdollisia ajatuksia käytöksen takana. Samalla tavoinhan kirjallisuuden lukija oppii ymmärtämään ongelmaisten roolihenkilöiden kaltaisia ajatuksia ja toimintaa.

Tärkeintä lähestymistavassa on se, miten tukiäly ottaa oppiaineistonsa suodattamiseen tarvitsemansa ohjeet ja arvot käyttäjältään, ei järjestelmän kehittäjältä tai regulaattorilta. Vaarana on luonnollisesti, että käyttäjä haluaa itselleen tai muille pahaa tai ei ikään kuin ymmärrä omia halujaan haitallisiksi, ja pyytää tekoälyltä apua vääriin päämääriin. Tämän ja isoveljen ohjaaman hyväntahtoisen tekoälyn ongelmat ja hyödyt voi asettaa vastakkain, ja arvioida, minkälaisessa yhteiskunnassa haluamme elää.

Kummallekin tulevaisuudelle yhteistä on yleisten kielimallien pyrkimys tuottamansa tekstin sisäiseen koherenssiin ja viitekehyksen mukaiseen mahdollisimman vähäiseen ristiriitaisuuteen valituksi tulleen opetusaineiston osan kanssa. Koherenssin ymmärtäminen on olennaista. GPT-3:n tapauksessa se ei näytä tarkoittavan virheetöntä päättelyä tai laskutaitoa. Puutteistaan huolimatta GPT-3:n kommentointi tyypillisissä kysymyksissä on usein koherentimpaa ja uskottavampaa kuin akateemistenkaan ihmisten keskustelu omien toimialojensa yleisistä asioista sosiaalisessa mediassa.

Olemme etenemässä kohti tulevaisuutta, jossa kuka tahansa voi keskustella tukiälyn kanssa Tukiäly ammentaa näkemyksensä kaikesta saatavilla olevasta kirjoitetusta tekstistä, kuvamateriaalista ja käyttäjänsä havaintoympäristöstä. Tukiäly on läsnä sekä työssä että arjessa, avartaa näkökulmiamme, parantaa ajattelumme koherenssia ja neuvoo uusiin taitoihin. Youtube toimii jo hieman tähän tapaan näyttäen kysyjälle lyhyitä opetusvideoita ratkaisuksi lähes mihin tahansa ongelmaan.

Työllistyminen, selviäminen elämässä, merkityksellisyyden tunne ja kansallinen menestys riippuvat kaikki erikoistuneesta osaamisesta, joka väistämättä ilman tukiälyn apua jää vajaaksi verrattuna niihin, jotka tukiälyn avun ottavat vastaan. Tämän vuoksi tukiälyn tulisi mahdollisuuksien mukaan olla käyttäjänsä ohjaama sen sijaan että se on yhteiskunnan tai yritysten välikäsi kansalaisten ja asiakasmassojen ohjaamiseen.

On joka tapauksessa täysin ilmeistä, että jo nykytasoisia kielimalleja voidaan käyttää sosiaalisessa mediassa, viestipalveluissa ja asiakaspalveluissa ihmisten kanssa keskusteluun sellaisin tavoin ja sellaisessa laajuudessa, että sitä voidaan pitää yhteiskunnille vaarallisena. Tämä ongelma vaatii ratkaisunsa, joka perustuu joko tekoälyn ja sille annettujen tavoitteiden läpinäkyvyyteen tai yhteiskunnan sisällölliseen kontrolliin.

Viestit päättäjille ja kansalaisille, valitkaa viisaasti:

Tekoälyn visioiden edessä on helppo pelästyä. Aiheellisia ja aiheettomia uhkakuvia on helppo maalailla ja uskoa, että tekoäly tulee harhauttamaan muita ihmisiä massoittain. Eurooppalaisiin arvoihin vedoten voidaan vaatia tekoälylle rajoitteita, jotka takaavat sen turvallisen käytön ja näin tulee pyrkiä tekemään. Tehtävä ei ole helppo.

Pelkojen ohjaamana kielletään helposti paljon sellaistakin, joka olisi arvokasta, jos siitä ei vielä ole kokemuksia eikä riittävän voimakkaita positiivisia visioita. Nyt esitelty komission ehdotus tekoälyn reguloinniksi (Artificial Intelligence Act) näyttää sivuuttavan tukiälynäkökulman. Kielletyiksi ehdotetut tekoälyn käytännöt voi helposti tulkita siten, että puhevikaisen puhetta korjaavan ja katsetta sekä muuta ulkonäköä korjaavasta tekoälystä on ilmoitettava katsojille, näkövammaisilta ja kuulovammaisilta kielletään tekoälyn apu kasvojen tai äänen tunnistamiseen ja tekoälyä oman ajattelun auttamiseen käyttävältä estetään tekoälyn roolitus, jos joku voisi käyttää sitä omaksi haitakseen.

Suomen tulisi muistaa eurooppalaisten arvojen lisäksi myös suomalaiset arvot – nopeana teknologian omaksujana, kekseliäänä, itsenäiseen ja muiden ihmisten kunnioitukseen perustuvaan teknologian käyttöön tottuneina, liian ahtaat yksilöille asetetut rajat eivät edesauta Suomen menestystä. Suomen tulisi nostaa tukiälyn näkökulma, itsenäiskäytön näkökulma ja muut kekseliään kansan positiiviset visiot esiin ja huolehtia siitä, etteivät ne hautaudu EU:n piirissä maalailtujen pelkojen alle. Rajoitukset on kohdistettava siten, että massamanipuloinnin ja massavalvonnan estäminen ei estä tukiälyn itsenäiskäyttöä vammaisten apuvälineissä ja itsenäisten ihmisten personoituina neuvonantajina. Tämä näkökulma puuttuu valtioneuvoston kirjelmästä (U 28/2021) ja valtioneuvoston selvityksestä (E 62/2021) komission tiedonantoon eurooppalaisesta lähestymistavasta tekoälyyn.

Suomen tulisi edellisen lisäksi ohjata EU:n toimintaa kielimalleihin liittyen siten, että eurooppalaisten kielien kyvykkyys ja eurooppalaisten kulttuurien tuntemus kielimalleissa on turvattu. On aivan ilmeistä, että esimerkiksi saatavilla oleva suomenkielinen julkaistu aineisto ei riitä kehittyneen keskustelevan tekoälyn opettamiseen. Kielimallit tullaan opettamaan kansainvälisesti saatavilla olevin aineistoin ja kukin kieli ja kulttuuri tulee voida lisätä niihin, jotta kielet ja kulttuurit eivät kuihdu.

EU:n tulisi tukea avoimen yhteisön tekoälyn kehitystä ja pyrkiä osana yleissivistävää koulutusta, elinikäistä oppimista, kirjastojen ja yleisradiopalvelun kaltaista kansalaisten tiedontarpeen tyydytystä käyttää merkittävä panos siihen, että minimissään GPT-3 -tasoinen personoitava tekoäly olisi kaikkien EU:n kansalaisten omilla kielillään saatavilla. Osaaminen ja ohjelmistot tähän ovat olemassa. Mittatikkuna todettakoon, että tämä olisi toteutettavissa koko EU:n laajuisena murto-osalla yleisradiotoiminnan budjettia, ja vaikutus kansalaisten osaamisen kehitykseen olisi radikaali.

Suomen tulisi rohkeasti ryhtyä edistämään tukiälyä EU:n laajuisena kansalaisoikeutena, ja se tulisi tehdä avoimen ja läpinäkyvän mallin mukaan siten, että kukin käyttäjä voi asettaa haluamansa arvot ja persoonan itseään auttavalle tekoälylle, siis sille omalle tukiälylleen, jonka kanssa asioitaan ja ajatuksiaan haluaa tarkastella.