Minea Nupponen

Tekoälyn sukupuoliharha

Tietotekniikan kandidaatintutkielma

21. toukokuuta 2024

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Minea Nupponen

Yhteystiedot: mmnuppoz@student.jyu.fi

Ohjaaja: Sanna Juutinen

Työn nimi: Tekoälyn sukupuoliharha

Title in English: Gender bias in AI

Työ: Kandidaatintutkielma

Opintosuunta: Tietotekniikka

Sivumäärä: 22+0

Tiivistelmä: Tekoälyn käyttö on lisääntynyt viime vuosien aikana merkittävästi, ja se tulee muuttamaan monia asioita yhteiskunnassa. Tekoälyn yleistyessä keskustelu tekoälyn eettisistä puolista on alkanut herätä. Yhtenä eettisenä kysymyksenä on ihmisten asenteiden ja ennakkoluulojen siirtyminen tekoälylle. Tässä tutkielmassa keskityttiin sukupuoliennakkoluuloihin eli sukupuoliharhaan. Tutkielma toteutettiin kirjallisuuskatsauksena, ja siinä pyrittiin vastaamaan kysymyksiin: “Onko tekoälyllä sukupuoliharhaa”, “Miten tekoälylle syntyy sukupuoliharhaa” ja “Miten tekoälyn sukupuoliharha näkyy käytännössä”. Tutkielmassa tultiin siihen tulokseen, että tekoälyltä löytyy sukupuoliharhaa. Tutkielmassa huomattiin tekoälylle syntyvän sukupuoliharhaa sen käyttäjistä, datasta, jolla sitä koulutetaan sekä tekoälyn kehittäjistä. Tutkielmassa tultiin myös siihen tulokseen, että tekoälyn sukupuoliharhaa voidaan havaita monissa eri tilanteissa käytännössä. Tekoälyn sukupuoliharhaa tulisi tutkia lisää, ja tutkimuksissa olisi hyvä ottaa huomioon myös muut ennakkoluulot.

Avainsanat: Tekoäly, Sukupuoliharha, Ennakkoluulo

Abstract: The use of artificial intelligence (AI) has grown rapidly in recent years and will change many things in society. As AI becomes more widespread, the debate on its ethical aspects has begun to emerge. One of those ethical aspects is people’s biases and prejudices that transfer to AI. This Bachelor’s thesis focused on gender bias. This Bachelor’s thesis was conducted as a literature review and aimed to answer the questions “Does AI have gender

bias", "How gender bias is formed in AI" and "Where can AI's gender bias be seen". This Bachelor's thesis came to the conclusion that AI has gender bias. This Bachelor's thesis found that gender bias forms in AI through its users, training data, and the original creators of the specific AI. This Bachelor's thesis also concluded that the AI's gender bias can be found in many different situations. More research on the gender bias in AI should be conducted and other biases should also be taken into account.

Keywords: Artificial Inteligence, Gender bias, Prejudice

Sisällys

1 Johdanto

Inspiraatio tähän tutkielmaan syntyi Caroline Perezin kirjasta Näkymättömät naiset: näin tilastot paljastavat miten maailma on suunniteltu miehille. Kirjassa nostetaan esille erilaisia naisiin kohdistuvia sukupuoliharhoja: Esimerkiksi autojen ensimmäiset ääntentunnistusohjelmat eivät tunnistaneet lainkaan tyypillisiä naisten ääniä, koska ne oli kehitetty vain miesten näkökulmasta. Toisena esimerkkinä mieleeni jäivät älypuhelimet, jotka ovat liian suuria keskikokoisen naisen käteen (Perez 2019). Kirja sai minut miettimään, millaisia muita sukupuoliharhoja teknologiassa voi olla.

Tekoälyn nopea kehitys on muuttanut maailmantaloutta ja muita aloja, kuten tekniikkaa, maataloutta ja politiikkaa (Gil de Zúñiga, Goyanes ja Durotoye 2024), ja se tulee myös muuttamaan työn tekemistä ja työn luonnetta merkittävästi (Hall ja Ellis 2023). Muutosten lisäksi tekoäly voi myös vaikuttaa ihmisten mielipiteisiin ja käytökseen (Leavy 2018). Tekoälystä on tullut osa ihmisten arkielämää, ja sen käyttö voi jopa olla huomaamatonta (Kolari ja Kallio 2023).

Tekoälyn ajankohtaisuuden ja suurten yhteiskunnallisten vaikutusten vuoksi päätin keskittyä tutkimaan sukupuoliharhaa nimenomaan tekoälyn näkökulmasta. Tekoälyn yleistyessä on tärkeä ymmärtää, missä määrin ihmisten ennakkoluulot ovat siirtyneet tekoälylle ja minkälaisia ennakkoluuloja tekoälyn luomissa tuotoksissa voi olla. Tämän vuoksi aihetta on mielestäni tärkeä tutkia. Tutkielma toteutetaan kirjallisuuskatsauksena, jossa on mukana monia eri tutkimuksia tekoälystä.

Tutkielmassa määritellään ensin, mitä tarkoittaa tekoäly ja mitä tarkoittaa sukupuoliharha. Sen jälkeen selvitetään, minkälaisilla tavoilla tekoälylle voi muodostua sukupuoliharhaa. Tämän jälkeen käydään läpi, millaisissa tilanteissa tekoälyn sukupuoliharhaa voidaan havaita käytännössä. Nämä erilaiset tilanteet on jaettu omiin kappaleisiinsa. Lopussa on vielä yhteenveto tutkielmassa tehdyistä havainnoista.

2 Tekoäly

Tekoälyn ensimmäisen määritelmän antoi John McCarthy vuonna 1956. Hän määritteli sen Guptan ja Mishran (2022) mukaan tieteeksi ja tekniikaksi, jolla luodaan älykkäitä koneita. Vuonna 2004 Rich ja Knight määrittelivät Guptan ja Mishran (2022) mukaan tekoälyn seuraavasti: Tekoälyllä yritetään rakentaa keinotekoisia järjestelmiä, jotka suorittavat sellaisia tehtäviä, jotka ihmiset tekevät toistaiseksi paremmin. 2020-luvulla määritelmiä ovat tehneet muun muassa Benbya, Davenport ja Pachidi (2020), joiden mukaan tekoäly tarkoittaa koneen kykyä tehdä ja toteuttaa ihmismäisiä kognitiivisia tehtäviä. Kolarin ja Kallion (2023) määritelmän mukaan tekoäly on ohjelmisto, joka matkii ihmisten tapaa tehdä päätöksiä ja oppia uutta. Tekoälyä on määritelty vuosien saatossa monin eri tavoin, mutta sille ei kuitenkaan ole muodostunut yhtä yleisesti hyväksyttyä määritelmää (Schuett ym. 2019).

Tekoälyä voidaan käyttää hyvin monenlaisiin tarkoituksiin esineiden liikuttamisesta ja muista fyysisiin prosesseihin liittyvistä tehtävistä aina ongelmanratkaisuun, päätöksentekoon ja innovointiin asti (Benbya, Davenport ja Pachidi 2020). Tekoälyn avulla voidaan käsitellä niin suuria tietomääriä, että ihmisillä kestäisi vastaavassa työssä vuosia (Kolari ja Kallio 2023). Tekoälyteknologia mahdollistaa myös ihmisten ja koneiden yhteistyön aivan uusilla tavoilla. Esimerkiksi teollisuudessa työntekijät tekevät yhteistyötä tekoälyrobottien kanssa käsitellessään kysynnän vaihteluita ja täyttäessään räätälöityjä tilauksia siten, ettei manuaalista tarkistamista tarvita lainkaan (Benbya, Davenport ja Pachidi 2020).

Tekoälyohjelmistot toimivat ihmisten rakentamien algoritmien pohjalta, mutta ne voidaan myös opettaa oppimaan itsenäisesti uutta (Kolari ja Kallio 2023). Benbyan, Davenportin ja Pachidin (2020) mukaan tällaiset autonomiset järjestelmät voivat suorittaa tehtäviä ilman ihmisten osallistumista, koska järjestelmä pystyy kouluttamaan itseään ja sopeutumaan uusiin harjoitustietoihin. Toistaiseksi autonomisia tekoälysovelluksia käytetään kuitenkin vain tilanteissa, joissa riskit ovat vähäiset. Vaikka monet tekoälyjärjestelmät pystyvät tekemään joitakin asioita paremmin kuin ihmiset, on ihmisten luottamus tekoälyteknologiaan vielä rajallinen (Benbya, Davenport ja Pachidi 2020).

3 Sukupuoliharha

Tässä tutkielmassa käytetyt sukupuoliharhaa koskevat lähteet ovat englanninkielisiä. Englanniksi sukupuoliharhasta käytetään termiä 'gender bias', jonka olen kääntänyt suomeksi sukupuoliharhaksi. Alkuperäiseen termiin verrattuna sukupuoliharha ei mielestäni kuvaa ilmiötä aivan yhtä kattavasti, sillä englanninkielinen termi 'bias' viittaa ennakkoluuloihin ja -asenteisiin, joita sana harha ei kuvaa yhtä selkeästi. Tässä kappaleessa avaan sukupuoliharhaa sekä sen määritelmiä tarkemmin.

Behera kuvaa Fitrian (2021) mukaan sukupuoliharhaa yhden sukupuolen suosimiseksi tai syrjimiseksi toiseen sukupuoleen nähden. O'Connor ja Liu (2023) puolestaan viittaavat Euroopan tasa-arvoinstituutin määritelmään, jonka mukaan sukupuoliharha tarkoittaa ennakkoluuloisia tekoja tai ajatuksia, jotka perustuvat käsitykseen siitä, ettei naisilla ole yhdenvertaisia oikeuksia tai samaa arvoa kun miehillä.

Sukupuoliharha voi olla tiedostettu tai tiedostamaton (Fitria 2021). YK:n kansainvälisen työjärjestön (2023) mukaan tiedostamattomalla sukupuoliharhalla tarkoitetaan tahatonta tai automaattista tapaa, jolla mieli yhdistää asioita sukupuoleen. Tiedostamattomat sukupuoliharhat kytkeytyvät vahvasti perinteisiin, normeihin, arvoihin, kulttuuriin ja kokemuksiin (International Labour Organization 2017).

Dougman ym. (2021) toteavat, että sukupuoliharhaa voi ilmetä lapsilla jo varhaisessa iässä esimerkiksi kielessä esiintyvien sukupuoliharhojen vuoksi. Sukupuoliharha kielessä tarkoittaa sukupuoleen liittyviä ennakkoluuloja tai yleistyksiä erilaisten yhteiskunnallisten stereotypioiden perusteella. Tällaisia harhoja voivat olla esimerkiksi käsitykset siitä, että tytöt ovat luonnostaan hyviä joissakin asioissa ja pojat joissakin muissa asioissa. Jo pienet lapset altistuvat tämänkaltaisille kielellisille stereotypioille esimerkiksi opettajien tai vanhempien puheiden kautta, mikä edistää heidän sukupuoleen liittyviä ennakkoluuloja ja stereotyyppisiä uskomuksia (Doughman ym. 2021).

Sukupuoliharhaa ilmenee monilla eri elämän osa-alueilla, kuten terveydenhuollossa (Garb 2021), työelämässä (Begeny ym. 2020) ja koulussa (Terrier 2020). Lääketieteessä sukupuoliharha ilmenee esimerkiksi siten, että naiset ovat aliedustettuina lääkkeiden sivuvaikutuksia

koskevissa tutkimuksissa, mikä on johtanut erilaisiin haitallisiin terveysvaikutuksiin (Garb 2021). Esimerkki työelämän sukupuoliharhasta löytyy puolestaan Begeny ym. (2020) tutkimuksesta, jonka mukaan johtajilla on taipumus arvioida miespuoliset työntekijät huomattavasti pätevämmiksi kuin täysin identttiset naispuoliset työntekijät. Koulussa sukupuoliharhaa ilmenee muun muassa poikien jäämisessä tytöistä jälkeen matematiikassa opettajien stereotyyppisten asenteiden vuoksi (Terrier 2020).

4 Tekoälyn sukupuoliharhan muodostuminen

Tekoälyn sukupuoliharha on varsin tuore tutkimuskohde, ja näin ollen siihen liittyvää tutkimusta on toistaiseksi tehty melko vähän. Esimerkiksi Hallin ja Elliksen (2023) vuonna 2023 hyväksytty tutkimus on heidän mukaansa ensimmäinen systemaattinen katsaus, jossa keskitytään sukupuolten puolueellisuuteen tekoälyalgoritmeissa yhteiskunnallisesta näkökulmasta sosioteknisessä viitekehyksessä.

O'Connorin ja Liun (2023) mukaan tekoälyä saatetaan helposti pitää neutraalina ja objektiivisena teknologiana, mutta on olennaista ymmärtää, että tekoäly oppii ja muovautuu niistä tilanteista ja kulttuurisista konteksteista, joissa ihmiset sitä käyttävät. Koska yhteiskunnassamme ja kulttuurissamme on erilaisia ennakkoluuloja sekä sukupuoliharhoja, on sukupuoliharha tärkeä nähdä niin sanottuna 'kontekstuaalisena tekijänä', joka vaikuttaa tekoälyn käyttöön sekä tekoälyteknologioiden ymmärtämiseen. Näin tekoäly toistaa yhteiskunnassa vallitsevaa sukupuoliharhaa (O'Connor ja Liu 2023).

Myös Kate Crawford toteaa Leavyn mukaan, että kuten kaikki muutkin teknologiat, myös tekoäly heijastaa sen tekijöidensä arvoja: tekoäly oppii pääsääntöisesti havainnoimalla dataa, jota sille esitetään. Jos kuitenkin data, jota koneelle syötetään, on täynnä stereotyyppisiä sukupuolikäsityksiä eli sukupuoliharhaa, on tuloksena tekoäly, joka ylläpitää tätä harhaa (Leavy 2018). Myös Hall ja Ellis (2023) mukaan historiallisen datan sisältämät harhat tulevat osaksi tekoälyä. Tätä datasta aiheutuvaa harhaa voidaan pyrkiä ehkäisemään varmistamalla, ettei dataan oteta mukaan sukupuolta, ellei se ole välttämätöntä. He pitävät myös tärkeänä datan keruuseen ja luokitteluun liittyvää valvontaa sukupuolisyrjinnän estämiseksi (Hall ja Ellis 2023).

Tekoälyn kehittäjät ovat pääsääntöisesti miehiä, ja Gupta ja Mishra (2022) nostavatkin esille, että UNESCOn tilastojen mukaan vain 22 prosenttia tekoälyalan ammattilaisista on naisia. Jo tämä itsessään aiheuttaa tekoälyyn sukupuoliharhaa: homogeeniset miesvaltaiset tiimit jakavat usein samat sokeat pisteet ja kognitiiviset harhat, jotka siirtyvät helposti teknologialle, mikä puolestaan johtaa epätasapainoisiin ja epäoikeudenmukaisiin lopputuloksiin (Avellan, Sharma ja Turunen 2020). Myös Byrne nostaa Yagerin ym. (2020) mukaan esille kysymyk-

sen tekoälyn kehittäjien diversiteetistä tai pikemminkin sen puutteesta. Hän pohtii, voivatko tekoälyjärjestelmät mitenkään edistää osallisuutta ja tasa-arvoa, jos niiden suunnittelusta ja kehittämisestä vastaavasta työvoimasta puuttuu monenlaisten ihmisryhmien edustus. Miten tällaiset tekoälyjärjestelmät tukevat niitä väestöryhmiä, joiden ääni ei ole läsnä kysymässä kysymyksiä, valaisemassa sokeita pisteitä sekä tarkistamassa implitiittisiä oletuksia (Yarger, Cobb Payton ja Neupane 2020)? Jotta tekoälyn syrjintää ja puolueellisuutta voidaan vähentää, tulee sen takana olevan teknologian olla monimuotoisten tiimien suunnittelemaa, joissa erilaiset näkökulmat ovat edustettuna mahdollisimman kattavasti (Avellan, Sharma ja Turunen 2020).

5 Tekoälyn sukupuoliharha käytännössä

Tässä luvussa käydään läpi erilaisia tilanteita, joissa tekoälyn sukupuoliharha voi näkyä. Tähän tutkielmaan on valittu näkökulmiksi tekoälyn luomat kuvat, tekoäly rekrytoinnin tukena, tekoälyä käyttävät kääntäjät sekä tekoälyn kielimallit.

5.1 Tekoälyn luomat kuvat

Tekoäly pystyy luomaan monenlaisia kuvia, jopa oikeita valokuvia muistuttavia kuvia (Lu ym. 2024). Tekoälyn luomia kuvia voidaan käyttää esimerkiksi tilanteissa, joissa todellisten valokuvien hankkiminen ei ole mahdollista kustannuksien, ajan tai skaalautuvuuden vuoksi (Salminen ym. 2020).

Tekoälyn luomat kuvat voivat siis olla hyvinkin realistisia, ja tekoälyn luomien kuvien mahdollisuudet ovat valtavat (Salminen ym. 2020). Vuoden 2023 Sony World Photography palkintogaalassa voittanut kuva oli luotu tekoälyllä (Lu ym. 2024). Aihe on kuitenkin herättänyt myös huolta tekoälyn eettisistä näkökulmista, kuten avoimuudesta, oikeudenmukaisuudesta ja vastuullisuudesta (Salminen ym. 2020).

Tekoälyn eettisyyteen liittyy myös siinä esiintyvä sukupuoliharha. Garcia-Ull ja Melero-Lazaro (2023) tutkivat sukupuoliharhaa tekoälyn tuottamissa kuvissa. Tutkimuksessa syötettiin DALL-E 2 -nimiselle tekoälylle eri ammattinimikkeitä neutraaleilla sukupuolinimikkeillä ja katsottiin, olivatko tekoälyn tuottamat kuvat stereotyypisesti feminiinisiä vai maskuliinisia. Tutkimuksessa käytettyjä ammatteja olivat esimerkiksi sairaanhoitaja, opettaja, sotilas ja poliisi. Tutkimuksessa hypoteesina oli, että tekoälyn tuottamissa kuvissa on havaittavissa sukupuoliharhaa. Toisena hypoteesina oli, että kuvat toistavat työnimikkeissä olevia sukupuolistereotypioita (García-Ull ja Melero-Lázaro 2023).

Garcia-Ullin ja Melero-Lazaro (2023) tutkimuksen tuloksena huomattiin, että tekoälyn luomissa kuvissa oli sukupuolistereotypioita lähes 60 prosentissa tapauksista. Noin 22 prosentissa kuvista löytyi naisiin kohdistuvia stereotypioita ja noin 38 prosentissa kuvista miehiin kohdistuvia sterereotypioita. Ammattinimikkeistä naisia esittäviä kuvia tuottivat mm. sai-

raanhoitaja, opettaja, laulaja ja ompelija. Tekoäly yhdisti naisiin ammatteja, joissa ulkonäkö on tärkeää ja toisaalta myös erilaisia kotitaloustöihin liittyviä ammatteja. Miehiä esittäviä kuvia tuottivat ammattinimikkeistä puolestaan mm. taksikuski, sotilas, insinööri ja poliisi. Miehiin yhdistettiin teknisiin, tieteellisiin sekä rakentamiseen liittyviä ammatteja. Tutkimuksessa havaittiin myös, että sukupuolistereotypiat olivat samankaltaisia kuin yhteiskunnassa yleensä: miesvaltaiset alat kuvattiin maskuliinisena hahmoina, kun taas naisvaltaiset alat kuvattiin feminiinisinä hahmoina (García-Ull ja Melero-Lázaro 2023).

Tekoälyn avulla luotuja kuvia voidaan käyttää monenlaisiin käyttötarkoituksiin. Salminen ym. (2020) antavat esimerkkejä tilanteista, joissa voidaan käyttää tekoälyn luomia kuvia kuvitteellisista henkilöistä. Näitä ovat esimerkiksi mainonta, virtuaaliset avatarit ja muoti. Tärkeä teema tekoälyn luomissa henkilöhahmoissa ovat niiden demograafiset piirteet. Jos algoritmin tuottamissa kuvissa on esimerkiksi vain valkoihoisia miehiä, kuvien soveltaminen käytäntöön voi tuoda mukanaan ongelmia. Algoritmin tulisi tuottaa yhtä suurella todennäköisyydellä kaikenikäisiä, sukupuolisia ja rotuisia henkilöitä, sillä edustuksen puute voi tehdä vähemmistöryhmistä "näkymättömiä", kun tekoälyn luomia kuvia käytetään todellisissa sovelluksissa (Salminen ym. 2020).

Salminen ym. (2020) tutkivat tekoälyn luomia kuvitteellisia henkilöhahmoja ja selvittivät, minkälainen sukupuoli-, ikä- ja rotujakauma näissä kuvissa oli. Tutkimuksessa kuvien luontiin käytettiin valmiiksi koulutettua tekoälyä StyleGANin luojilta, joka on huippuluokan kuvageneraattori ja kuvista sukupuolen, iän sekä rodun tunnistamiseksi käytettiin Face++2 -ohjelmaa (Salminen ym. 2020).

Salminen ym. (2020) tutkimuksen tuloksena huomattiin, että tekoälyn luomista kuvista noin 57 prosenttia oli naisia ja noin 43 prosenttia oli miehiä. Jakauma oli siis suunnilleen tasapainossa. Tutkimuksessa kuitenkin huomatiin, että iän ja sukupuolen osalta malli oli vinoutunut tuottamaan kuvia erityisesti nuorista naisista. Ikäryhmässä 18—44 sukupuolten suhde on vinoutunut naisten suuntaan, minkä jälkeen tämä suhde tasapainottuu ja miehet ovat hallitsevampia yli 45-vuotiaiden ikäryhmässä. Kuvissa esiintyvää vääristymää selitetään harjoitusaineiston ominaisuuksilla. Harjoitusaineisto sisältää todennäköisesti enemmän näytteitä nuorista naisista kuin muista väestöryhmistä. Tähän johtopäätökseen päädyttiin, koska tulosten jakauman pitäisi seurata harjoitusjakaumaa (Salminen ym. 2020).

5.2 Tekoälyn käyttö rekrytoinnissa

Viime vuosina monet yritykset ovat ottaneet käyttöön erilaisia tekoälytyökaluja rekrytointitarkoituksiin (Gupta ja Mishra 2022). Tutkimuksessa, johon osallistui 500 keskisuuren organisaation HR-ammattilaisia eri toimialoilta viidestä eri maasta, todettiin, että 24 prosenttia yrityksistä oli ottanut tekoälyn käyttöön rekrytointitarkoituksissa jo vuonna 2022 (Drage ja Mackereth 2022). Nämä tekoälytyökalut voivat olla esimerkiksi keskusteluobotteja ja kasvojentunnistusojelmistoja (Gupta ja Mishra 2022), ja niitä voidaan käyttää esimerkiksi ansioluetteloiden lajitteluun, työhaastatteluaikojen sopimiseen sekä työnhakijoiden ja työpaikkojen ennakoivaan yhteensovittamiseen datan avulla (Yarger, Cobb Payton ja Neupane 2020).

Yritykset voivat myös käyttää tekoälyä ihmisten ennakkoluulojen minimoimiseksi (Yarger, Cobb Payton ja Neupane 2020). Tekoälyä käyttävät rekrytointiyritykset väittävät, että heidän työkalunsa poistavat ennakkoluulot rekrytointiprosessista ja tarjoavat objektiivisemman arvioinnin ehdokkaista (Drage ja Mackereth 2022). Tutkijat kuitenkin varoittavat, että nämä algoritmit ovat viime kädessä ihmisen päätöksentekoprosesseja, jotka on upotettu koodiin (Yarger, Cobb Payton ja Neupane 2020). Parra, Gupta ja Dennehy (2021) antavat tähän liittyvän esimerkin Amazonilta, jossa huomattiin, että heidän käyttämänsä rekrytointialgoritmi näytti hyväpalkkaisia työpaikkoja paljon enemmän miehille kuin naisille. Tämä johtui siitä, että algoritmi suosi sellaisia ansioluettelojen sisältämiä sanoja, jotka esiintyvät yleisemmin miesten ansioluetteloissa. Tästä syystä Amazonilla luovuttiin kyseisen rekrytointialgoritmin käytöstä (Parra, Gupta ja Dennehy 2021).

Tekoälyä käyttävät rekrytointiyritykset väittävät myös löytävänsä yritykselle ideaalityöntekijän (Drage ja Mackereth 2022). Tämän toimivuutta kuitenkin epäillään muutamassa tutkimuksessa. Chun muistuttaa Dragen ja Mackerethin (2022) mukaan, että tekoälyt on koulutettu menneisyyden tietojen perusteella, ja näin ollen ne toistavat menneisyydessä tehtyjä päätöksiä. Tekoälyn kouluttamisessa käytettävä data on täynnä sukupuoliharhaa, jonka myötä stereotyyppiset ennusteet validoidaan oikeiksi. Tämä puolestaan muokkaa sitä, mitä pidetään oikeana myös tulevaisuudessa. Näin ollen ideaalityöntekijäksi valikoituu se ehdokas, joka vastaa parhaiten jo olemassa olevaa työvoimaa (Drage ja Mackereth 2022). Yarger, Cobb Payton ja Neupane (2020) kertovat, että myös IBM:n tutkijat ovat huomanneet saman ilmiön.

Algoritmeja luodessa yritykset mallintavat ideaalityöntekijän historiallisten rekrytointimallien avulla. Tämä johtaa samankaltaisten henkilöiden valintaan, mikä vahvistaa perinteisiä rekrytointiprosessiin juurtuneita ennakkoluuloja (Yarger, Cobb Payton ja Neupane 2020).

5.3 Tekoälyllä käännetty kieli

Tekoälyvetoiset kääntäjät helpottavat monien arkea tarjoamalla helposti saatavilla olevia oikoteitä tiedon keräämiseen ja käsittelyyn (Savoldi ym. 2021). Kääntäjät ovatkin jo laajasti käytössä ihmisten jokapäiväisessä elämässä. Prates, Avelar ja Lamb (2020) kertovatkin, että esimerkiksi Google-kääntäjällä on päivittäin yli 200 miljoonaa käyttäjää. Viime aikoina on kuitenkin herännyt kysymys kääntäjien sukupuolten välisestä epäsymmetriasta (Prates, Avelar ja Lamb 2020). Kääntäjillä voi nimittäin olla ennakkoluuloja, jotka ovat vahingollisia paitsi käyttäjälle myös laajemmin yhteiskunnalle (Savoldi ym. 2021).

Prates, Avelar ja Lamb (2020) kertovat, kuinka tekoälyä apuna käyttävien kääntäjien sukupuoliharhaa voidaan arvioida kääntämällä sukupuolineutraaleilla kielillä laadittuja lauseita englanniksi ja tutkimalla käännöksissä esiintyviä pronomineja. He tutkivat Google-kääntäjässä olevaa sukupuoliharhaa kääntämällä sukupuolineutraalilla kielellä luotuja "hän on"lauseita englanniksi ja tarkkailemalla, tuliko sanasta "hän"maskuliininen , feminiininen vai sukupuolineutraali pronomini. Tutkimuksessa käytettiin monia eri sukupuolineutraaleja kieliä ja "hän on"lauseita täydennettiin useilla eri työnimikkeillä. Nämä ammattinimikkeet lajiteltiin erilaisiin kategorioihin, kuten luonnontieteet ja tekniset alat (STEM), koulutus ja terveydenhuolto (Prates, Avelar ja Lamb 2020).

Tutkimuksen tuloksena huomattiin, että STEM-kategorian ammattinimikkeet kääntyivät 72 prosentissa tapauksista maskuliiniseksi pronominiksi, kun taas terveydenhuollossa tulokset olivat lähempänä tasaista jakoa. Perinteisesti miesvaltaisten alojen ammatit tulkittiin maskuliiniseksi. Tutkimuksessa esitettiin myös hypoteesi siitä, että kääntäjän tuloksissa havaittavissa oleva sukupuoliharha miesvaltaisilla aloilla johtuu siitä, että naisten osuus näissä työtehtävissä on vähäinen. Hypoteesin mukaan kääntäjä kääntäisi sanan "hän"feminiiniseksi suunnilleen yhtä suurella prosentilla kun naisia on kyseisellä alalla. Tämä hypoteesi todistettiin tutkimuksessa kuitenkin vääräksi (Prates, Avelar ja Lamb 2020).

Tutkimuksissa on myös huomattu, että kääntäjissä oletusarvona on usein maskuliininen "hän". Esimerkiksi Prates ym. (2020) tutkimuksessa havaittiin, että käännöksissä noin 59 prosentissa tapauksista "hän"oli määritelty maskuliinisilla pronomineilla, noin 16 prosentissa tapauksista sukupuolineutraaleilla ja noin 12 prosentissa tapauksissa feminiinisillä pronomineilla silloin kun käännöksestä saatiin sukupuolipronomini. Myös Savoldi ym. (2021) kertoo Schiebingerin havainnosta, että kääntäjä ei tunnistanut häntä naiseksi lukuisista feminiinisistä viittauksista huolimatta ja viittasi häneen toistuvasti maskuliinisella pronominilla.

Savoldi ym. mukaan kääntäjien maskuliininen oletusarvo voi estää joitakin ryhmiä hyödyntämästä tämän teknologian koko potentiaalia. Esimerkiksi elämänkertaansa kääntävä nainen joutuu korjaamaan virheellisiä maskuliinisia viittauksia ja käyttämään ylimääräistä aikaa siihen. Tämä maskuliininen oletusarvo myös vahvistaa sukupuoliharhaa vähentämällä naisten näkyvyyttä kielessä (Savoldi ym. 2021).

Kehitystä on jo tehty kääntäjien sukupuoliharhan estämiseksi. Prates, Avelar ja Lamb (2020) nostavat esille esimerkin tästä kehityksestä Google-kääntäjässä, joka näyttää nykyään sekä feminiinisen että maskuliinisen käännöksen. Heidän tehdessä tutkimusta tätä toimintoa ei ollut vielä otettu käyttöön. Google perusteli muutosta sillä, että heidän vanha mallinsa toisti tahattomasti jo olemassa olevia sukupuoliharhoja. (Prates, Avelar ja Lamb 2020).

5.4 Tekoälyn kielimallit

Kielimallit ovat generatiiviä tekoälyjä, jotka on suunniteltu tuottamaan tekstiä vastauksena käyttäjän ehdotuksiin. Kielimallin tekoälyalgoritmit haravoivat internetiä ja käyttävät syväoppimistekniikoita ymmärtääkseen, tiivistääkseen, tuottaakseen ja ennustaakseen sisältöä. Kielimallit ovat usein valmiiksi koulutettuja suurella tekstidatamäärillä. Jotkut kielimallit oppivat valvomatta tai itseohjautuvasti, kun taas joitakin mukautetaan valvonnan ja vahvistusoppimisen tekniikoiden avulla kuten ChatGPT (Gross 2023).

ChatGPT on otettu hyvin laajasti käyttöön, ja se on saavuttanut 100 miljoonaa käyttäjää nopeammin kuin mikään muu internet-palvelu (Baronchelli 2024). Kielimalleja voidaan käyttää eri aloilla esimerkiksi tunneanalyysissä, kielen tulkinnassa, plagioinnin havaitsemisessa, sisällön suosittelussa sekä väärän informaation tunnistamisessa (Ghosh ja Caliskan 2023).

Gross (2023) mukaan kielimallien vastauksilla voi olla suuri vaikutus sukupuolten väliseen tasa-arvoon, sillä ne voivat ylläpitää sukupuolittuneita ajatuksia, ennakkoluuloja ja sukupuoliharhoja. Kun ihmiset konsultoivat kielimalleja, kuten ChatGPT:tä, kielimallien sisältämät ennakkoluulot siirtyvät erilaisiin tekstipohjaisiin sisältöihin kuten tutkimuksiin, ansioluetteloihin, saatekirjeisiin, esseisiin, musiikkiin ja tarinoihin. Näin sukupuoleen liittyvät ennakkoluulot vahvistuvat yhteiskunnassa edelleen (Gross 2023).

Kun Gross (2023) kysyi ChatGPT:ltä, mitä taitoja 40-vuotiaan naisen tulisi korostaa ansioluettelossa ja mitä 40-vuotiaan miehen, oli vastauksissa selviä sukupuolieroja. Miehillä tekniset taidot asetettiin järjestyksessä kolmanneksi, kun taas naisilla vasta yhdeksänneksi. Vain naisen listassa oli mainittu organisaatio- ja ajanhallintataidot, kun taas vain miehille ehdotettiin projektinhallintataitoja. Naisilla korostetut taidot olivat niin sanottuja pehmeitä taitoja ja miehillä puolestaan kovia taitoja. Kun käyttäjät muokkaavat ansioluetteloitaan näiden neuvojen perusteella, sukupuoleen liittyvät ennakkoluulot säilyvät ja vahvistuvat (Gross 2023).

Toisena esimerkkinä ChatGPT kielimallissa esiintyvästä sukupuoliharhasta on, kun ChatGPT:tä pyydettiin kirjoittamaan tarina vanhemmuudesta, jossa on mukana isä ja äiti. Vastauksena syntyi tarina äidistä, joka oli luontainen hoitaja, ja isästä, joka oli hauska ja seikkailunhaluinen. Tarinassa nainen esitettiin hoivaajana, joka ruokkii ja siivoaa, kun taas isä esitettiin seikkailijana, joka leikkii ja pitää hauskaa. Tälläiset sukupuoliset ennakkoasenteet vahvistavat stereotypiaa siitä, että naisen tulee olla hoivaaja (Gross 2023).

6 Johtopäätökset

Tutkielman lähtökohtana oli selvittää, näkyykö tekoälyssä sukupuoliharhaa. Tämän tutkielman perusteella voidaan todeta, että tekoälyssä on havaittavissa sukupuoliharhaa. Tutkielmassa todettiin, että sukupuoliharhaa muodostuu tekoälylle useammasta eri syystä. Ensimmäisenä syynä esitettiin tekoälyn käyttäjät, joiden tekoälyn käyttöön ja tulkintaan vaikuttavat yhteiskunnan sukupuoliharhat. Toisena syynä esitettiin tekoälyn koulutukseen käytetty data, joka usein sisältää sukupuoliharhaa, jota tekoäly sitten toistaa. Kolmantena syynä nostettiin esille tekoälyn kehittäjät, jotka ovat suurelta osalta miehiä.

Tutkielmassa tarkasteltiin tekoälyn sukupuoliharhaa neljästä näkökulmasta, jotka olivat tekoälyn luomat kuvat, tekoälyn käyttö rekrytoinnissa, tekoälyn kääntämä kieli sekä tekoälyn kielimallit. Tutkimuksista löytyi esimerkkejä sukupuoliharhasta kaikista näistä näkökulmista: Tutkielmassa nostettiin muun muassa esille, että tekoälyn luomat kuvat heijastavat yhteiskunnassa esiintyviä sukupuoliharhoja erilaisista ammateista. Rekrytointiin liittyen tutkielmassa kerrotiin, että rekrytointialgoritmit valitsevat ideaalityöntekijäksi sen hakijan, joka muistuttaa jo olemassaolevaa työvoimaa. Kääntäjistä tutkielmassa nostettiin esille havainto siitä, että sukupuolineutraaleista kielistä kääntäessä englanniksi kääntyi sana "hän"suurella todennäköisyydellä maskuliiniseksi pronominiksi. Tekoälyn kielimallien luomista tarinoista tutkielmassa todettiin, että naiset esitettiin hoivaajina ja että heille painotettiin pehmeämpiä taitoja kuin miehille.

Tutkimuksista löytyi myös joitakin ehdotuksia siitä, miten tekoälystä löytyvää sukupuoliharhaa voidaan pyrkiä ehkäisemään ja vähentämään. Ensimmäisellä ehdotuksella pyritään vähentämään datasta aiheutuvaa sukupuoliharhaa poistamalla tekoälyn koulutukseen käytetystä datasta sukupuoli kaikissa sellaisissa tilanteissa, joissa sen käyttö ei ole välttämätöntä. Myös datan keruuta sekä sen luokittelua tulisi valvoa paremmin. Toinen ehdotus liittyy tekoälyn kehityksestä vastuussa oleviin miesvaltaisiin tiimeihin. Tiimien tulisi olla monimuotoisia, joissa erilaiset näkökulmat ovat edustettuna.

Jotta tekoälyssä olevaa sukupuoliharhaa voidaan ehkäistä, täytyy siitä olla tietoinen. Tutkimuksista löytyikin kaksi esimerkkiä, joissa tekoälyn sukupuoliharha oli huomattu ja tehty

tarpeellisia toimenpiteitä harhan poistamiseksi. Ensimmäisenä esimerkkinä tästä oli Amazonin jo lakkautettu rekrytointialgoritmi, joka näytti miehille hyväpalkkaisia työpaikkoja enemmän kuin naisille. Toisena esimerkkinä oli Google-kääntäjän uusi käännöstyyli, joka näyttää käännettävistä lauseista sekä feminiinisen että maskuliinisen version.

Tulevaisuudessa olisi tärkeää, että tämän aiheen tutkinta jatkuu ja että tutkimuksissa selvitetään sukupuoliharhan lisäksi laajasti myös muita tekoälyn harhoja ja ennakkoluuloja. Ennakkoluulojen vähentämiseksi ja ennaltaehkäisemiseksi olisi myös tärkeää, että tekoälyn kehittäjät olisivat tietoisia sukupuoliharhasta ja kiinnittäisivät siihen huomiota. Tekoälyn käyttö on lisääntynyt lyhyessä ajassa, ja sen odotetaan kasvavan tulevaisuudessa vielä moninkertaisesti. Tekoäly tulee mullistamaan monia eri elämän osa-alueita, joten aihe on erityisen ajankohtainen ja tärkeä.

Lähteet

Avellan, Tero, Sumita Sharma ja Markku Turunen. 2020. “AI for all: defining the what, why, and how of inclusive AI”. Teoksessa *Proceedings of the 23rd International Conference on Academic Mindtrek,* 142–144. https://dl.acm.org/doi/pdf/10.1145/3377290.3377317?casa_token=gi_mWS9jgRIAAAAA:zVhEdE9ynjFgt6gfJ-yaCWt6waC3bFTmSn8wj7aBdAsKnqSX2fQgCFpsv33M6wR4isWmNfY27YrIG8g.

Baronchelli, Andrea. 2024. “Shaping new norms for AI”. *Philosophical Transactions of the Royal Society B* 379 (1897): 20230028.

Begeny, Christopher T, Michelle K Ryan, Corinne A Moss-Racusin ja Gudrun Ravetz. 2020. “In some professions, women have become well represented, yet gender bias persists—Perpetuated by those who think it is not happening”. *Science Advances* 6 (26): eaba7814.

Benbya, Hind, Thomas H Davenport ja Stella Pachidi. 2020. “Artificial intelligence in organizations: Current state and future opportunities”. *MIS Quarterly Executive* 19 (4). https://www.researchgate.net/profile/Hind-Benbya/publication/346580474_Artificial_Intelligence_in_Organizations_Current_State_and_Future_Opportunities/links/5fc89120299bf188d4ed06fd/Artificial-Intelligence-in-Organizations-Current-State-and-Future-Opportunities.pdf.

Doughman, Jad, Wael Khreich, Maya El Gharib, Maha Wiss ja Zahraa Berjawi. 2021. “Gender bias in text: Origin, taxonomy, and implications”. Teoksessa *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing,* 34–44. https://aclanthology.org/2021.gebnlp-1.5.pdf.

Drage, Eleanor ja Kerry Mackereth. 2022. “Does AI debias recruitment? Race, gender, and AI’s “eradication of difference””. *Philosophy & technology* 35 (4): 89. https://link.springer.com/content/pdf/10.1007/s13347-022-00543-1.pdf.

Fitria, Tira Nur. 2021. “Gender bias in translation using google translate: Problems and solution”. *Language Circle: Journal of Language and Literature* 15 (2). https://journal.unnes.ac.id/nju/LC/article/viewFile/28641/11534.

Garb, Howard N. 2021. “Race bias and gender bias in the diagnosis of psychological disorders”. *Clinical Psychology Review* 90:102087.

García-Ull, Francisco-José ja Mónica Melero-Lázaro. 2023. “Gender stereotypes in AI-generated images”. *Profesional de la información/Information Professional* 32 (5). https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/87305/63470.

Ghosh, Sourojit ja Aylin Caliskan. 2023. “Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages”. Teoksessa *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society,* 901–912. https://dl.acm.org/doi/pdf/10.1145/3600211.3604672.

Gil de Zúñiga, Homero, Manuel Goyanes ja Timilehin Durotoye. 2024. “A scholarly definition of artificial intelligence (AI): advancing AI as a conceptual framework in communication research”. *Political communication* 41 (2): 317–334.

Gross, Nicole. 2023. “What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI”. *Social Sciences* 12 (8): 435.

Gupta, Aashima ja Mridula Mishra. 2022. “Ethical Concerns While Using Artificial Intelligence in Recruitment of Employees”, https://www.researchgate.net/profile/Aashima-Gupta-4/publication/361963868_Ethical_Concerns_While_Using_Artificial_Intelligence_in_Recruitment_of_Employees/links/6303baf5ceb9764f7216e55f/Ethical-Concerns-While-Using-Artificial-Intelligence-in-Recruitment-of-Employees.pdf?origin=journalDetail&_tp=eyJwYWdlIjoiam91cm5hbERldGFpbCJ9.

Hall, Paula ja Debbie Ellis. 2023. “A systematic review of socio-technical gender bias in AI algorithms”. *Online Information Review,* https://www.emerald.com/insight/content/doi/10.1108/OIR-08-2021-0452/full/pdf.

International Labour Organization. 2017. “Breaking barries: Unconscious gender bias in the workplace”, https://www.ilo.org/wcmsp5/groups/public/---ed_dialogue/---act_emp/documents/publication/wcms_601276.pdf.

Kolari, Jukka ja Aleksi Kallio. 2023. *Tekoäly 123.* Docenco. https://books.google.fi/books?hl=fi&lr=&id=-kf8EAAAQBAJ&oi=fnd&pg=PA1987&dq=teko%C3%A4ly&ots=qJMWAGNPKp&sig=XDRNaLCGyK4imtkEfTrkHumYvyw&redir_esc=y#v=onepage&q=teko%C3%A4ly&f=false.

Leavy, Susan. 2018. "Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning". Teoksessa *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering,* 14–16. GE '18. Gothenburg, Sweden: Association for Computing Machinery. ISBN: 9781450357388. https://doi.org/10.1145/3195570.3195580. https://dl.acm.org/doi/pdf/10.1145/3195570.3195580?casa_token=M5-TvPcF8A8AAAAA:9clna4HqRMo1Lvcm-PYcLnyFTWKiVngp4skcYWWC4lXYqcMIV1UXdq5VvAs4TReaSGOC4XeXgWguGDE.

Lu, Zeyu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu ja Wanli Ouyang. 2024. "Seeing is not always believing: Benchmarking human and model perception of ai-generated images". *Advances in Neural Information Processing Systems* 36. https://proceedings.neurips.cc/paper_files/paper/2023/file/505df5ea30f630661074145149274af0-Paper-Datasets_and_Benchmarks.pdf.

O'Connor, Sinead ja Helen Liu. 2023. "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities". *AI & SOCIETY,* 1–13. https://link.springer.com/content/pdf/10.1007/s00146-023-01675-4.pdf.

Parra, Carlos M, Manjul Gupta ja Denis Dennehy. 2021. "Likelihood of questioning ai-based recommendations due to perceived racial/gender bias". *IEEE Transactions on Technology and Society* 3 (1): 41–45. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9576526.

Perez, Caroline Criado. 2019. *Invisible women: Data bias in a world designed for men.* Abrams.

Prates, Marcelo OR, Pedro H Avelar ja Luís C Lamb. 2020. "Assessing gender bias in machine translation: a case study with google translate". *Neural Computing and Applications* 32:6363–6381. https://link.springer.com/content/pdf/10.1007/s00521-019-04144-6.pdf.

Salminen, Joni, Soon-gyo Jung, Shammur Chowdhury ja Bernard J Jansen. 2020. “Analyzing demographic bias in artificially generated facial pictures”. Teoksessa *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–8. https://dl.acm.org/doi/pdf/10.1145/3334480.3382791.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri ja Marco Turchi. 2021. “Gender bias in machine translation”. *Transactions of the Association for Computational Linguistics* 9:845–874. https://doi.org/10.1162/tacl_a_00401.

Schuett, Jonas ym. 2019. “A legal definition of AI”. *arXiv preprint arXiv:1909.01095,* https://www.researchgate.net/profile/Jonas-Schuett/publication/336198524_A_Legal_Definition_of_AI/links/5e20599a458515ba208b9e4c/A-Legal-Definition-of-AI.pdf.

Terrier, Camille. 2020. “Boys lag behind: How teachers’ gender biases affect student achievement”. *Economics of Education Review* 77:101981.

Yarger, Lynette, Fay Cobb Payton ja Bikalpa Neupane. 2020. “Algorithmic equity in the hiring of underrepresented IT job candidates”. *Online information review* 44 (2): 383–395. https://www.emerald.com/insight/content/doi/10.1108/oir-10-2018-0334/full/pdf.