Kun tekoäly hallusinoi

30.5.2024

Puhe tekoälystä (Artificial Intelligence, AI) on täyttänyt kaikenlaista mediatilaa tehokkaasti etenkin sen jälkeen, kun OpenAI julkaisi oppiviin suuriin kielimalleihin (LLM, Large Language Models) perustuvan ChatGPT3.5-tekoälypalvelun marraskuussa 2022. Tekoälypohjaisia palveluja olemme kuitenkin käyttäneet sujuvasti jo vuosien ajan monenlaisten sovellusten osina, vaikkemme sitä aina ole tiedostaneetkaan. Verkon hakuohjelmat, keskustelubotit, somen suosittelutoiminnot, tekstiohjelman kielenhuolto- ja oikoluku tai Sirin ja Alexan kaltaiset tekoälyapurit ovat osa tätä kasvavan tekoälyavusteista arkeamme, johon teknologia kytkeytyy intiimein ja koko ajan kasvavin tavoin (AI-apureista ks. lisää Komonen 2024a). Tekoälypalvelut menestyvät jo ylioppilaskirjoituksissa (Pölönen 2024), suoriutuvat erinomaisesti älykkyystesteistä (Wasilewski & Jablonski 2024), ja sovellusten tarjoamat tulokset ällistyttävät yhä useammilla elämänalueilla. Olemme siirtymässä tekoälytodellisuuteen, siis todellisuuteen, johon erilaiset tekoälysovellukset vaikuttavat kasvavin tavoin.

Tämän tekstin kirjoittaminen on epäkiitollista puuhaa. Päivittäin julkaistaan uusia AI-sovelluksia ja aiempien sovellusten kehitysversioita sekä näiden mobiiliversioita samalla, kun tekoälysovelluksia integroidaan jo käytössä olevien ohjelmistojen uusiin versioihin, vaikkapa laajasti käytettyyn Microsoftin Office-tuoteperheeseen (ks. myös Komonen 2024b). Samaan aikaan tehdään jatkuvasti uudenlaisia oivalluksia tekoälyalgoritmien ominaisuuksista ja mahdollisuuksista. Tämän päivän havainnot vanhenevat ennätysmäisen nopeasti. ChatGPT-versioiden, Geminin ja Copilotin kaltaisten kielimallisovellusten rinnalle on internettiin ja muihin tietojärjestelmiin ehtinyt jo muodostua satojen, jopa tuhansien erilaisten tekoäly- eli AI-sovellusten digitaalinen ekosysteemi.

Valtiot, yritykset, organisaatiot, oppilaitokset ja yksityiset kansalaiset yrittävät nyt kukin tahoillaan ymmärtää, mitä kaikkea tämä voi tarkoittaa itselle ja omalle organisaatiolle: uhka ja mahdollisuus? Voiko tekoälyyn luottaa? Jälkimmäinen kysymys on tekoälysovellusten eettiseen käyttöön liittyvä ydinteema. Tarkastelen aihetta tässä artikkelissa humanistis-

yhteiskuntatieteelliseen taustaani nojaten lähinnä kulttuuriantropologin ja ammattikorkeakoulun media-alan yliopettajan näkökulmista.

Mitä saa, kun tilaa?

Kysymys tekoälyyn luottamisesta on aiheellinen, sillä erilaisten AI-palvelujen on yleisesti havaittu tarjoavan virheellistä sisältöä, vaikkakin houkuttelevan uskottavasti muotoiltuna: tekoäly *hallusinoi*, kuten nopeasti vakiintunut puhetapa kuuluu (Keary 2024; Leffer 2024; Maleki, Badmanabhan & Dutta 2024; vrt. Puttonen 2024). Kun promptaus eli käskyttäminen tuottaa paikkansapitämättömiä sisältöjä, puhe AI-hallusinaatioista piilottaa sanavalinnan taakse kätkeytyvän pyrkimyksemme inhimillistää tekoälyä ja tehdä siitä oman inhimillisen tietojärjestelmämme kanssa yhteismitallisen, ikään kuin sähköisen ihmisenkaltaisen kumppanin. Samaan viittaa puhe ”älykkäistä” järjestelmistä, palveluista tai sovelluksista: teknologiaa inhimillistetään, jotta ymmärtäisimme paremmin sen toimintaperiaatteita ja osaisimme mukautua paremmin sen kanssa toimimiseen (ks. myös Ruckenstein 2022). Kuitenkaan psykiatrit eivät ole ihastuneita hallusinaatio-sanan käyttöön tekoälysovellusten yhteydessä, koska sillä on oma diagnostinen luokittelunsa ihmisten hoidossa. Psykiatriassa hallusinaatiolla tarkoitetaan aistiharhaa, joka koetaan ilman ulkoista ärsykettä (Huttunen 2018). AI-hallusinaation sijaan ehdotetaan *satuilua* ja *vääristelyä* (Edwards 2023; Emsley 2023).

Olipa käsitteenä hallusinaatio, satuilu tai vääristely, sen viittauskohde on kuitenkin sama: promptaus eli tekoälysovellukselle annettu komento voi tuottaa paikkansapitämätöntä tietoa. Kuinka tällaisessa tilanteessa voimme varmistua siitä, mikä osa saadusta tuloksesta pitää paikkansa tai mihin tiedot perustuvat? Emme oikein mitenkään. Emme tiedä, millaisilla aineistoilla palvelun algoritmit on opetettu, millaisia vinoumia opetettuun aineistoon liittyy tai millaisia painotuksia eri näkökohdat saavat algoritmien toiminnassa. Tieteellisen luotettavuuden kannalta käsillä on perustavanlaatuinen ontologinen ja epistemologinen ongelma: millaisena promptauksen tuottama sisältö tulisi ylipäätään ymmärtää ja kuinka sitä tulisi arvioida totuus- ja todellisuusväittämien kannalta? Syntetisoivan ja läpinäkymättömän luonteensa vuoksi promptauksen palauttama tulos on tieteellisen tiedon lähteenä ongelmallinen verrattuna ”perinteisiin”

tutkimuksiin ja tutkimustuloksiin, joiden oikeellisuudesta kantavat vastuuta nimetyt tekijät nimettyine aineistoineen, tietolähteineen ja artikuloituine tietoteorioineen. Muut tutkijat voivat tällöin varmistua tulosten oikeellisuudesta itsenäisesti. Tämä ei onnistu AI-tulosten kohdalla – varsinkin, jos sama promptaus tuottaa eri kerroilla erilaisia tuloksia.

Mustan laatikon (black box) käsitteellä kuvataan sitä, mitä tekoälyalgoritmien toiminnassa varsinaisesti tapahtuu: vaikea sanoa, mitä. Edes tekoälypalveluja kehittävät tutkijat eivät osaa täsmällisesti selittää, miten algoritmit lopulta toimivat, kun ne reagoivat annettuihin komentoihin (Perriggo 2023). Kriitikot ovat jopa verranneet tekoälyn toimintaa magiaan, jossa promptaukset toimivat loitsuina (Gold 2023). Ajatus magiasta tulee kieltämättä joskus mieleen, kun sovellukset tarjoavat tuloksiksi pitkälle vietyjä jäsennyksiä monimutkaisista aineistoista, taidokkaan näköisiä kuva- ja videoesityksiä, kehittynyttä ohjelmakoodia tai muuta inhimillistä jäsennyskykyämme haastavaa sisältöä – vain muutaman lauseen kehotteella.

Tekoälysovelluksia käytetään jo laajasti vaikkapa eri tieteissä, teollisuudessa ja tietoverkkoympäristöissä, ja tämä kehitys etenee nopeasti yhä uusille alueille. Samalla rakentuu mielikuva tekoälystä jonakin uutena ja vallankumouksellisena tiedon tuottamisen välineenä, jolta odotetaan melkeinpä mitä vain ihmiskunnan valtaansa kaappaavasta singulariteetista ja sen ytimessä olevasta *yleisestä tekoälystä* (Artificial General Intelligence, AGI) erilaisten yhteiskunnallisten käytännön ongelmien arkisiin ratkaisuihin. Mistä on varsinaisesti kyse? Antropologi ja kielitieteilijä Ilana Gherson muistuttaa, että mikään koneen tuottama sisältö ei viittaa maailmaan eli sillä ei ole referenttiä eikä myöskään kontekstia todellisuudessa. Sisältö vain synnyttää sellaisen vaikutelman, koska algoritmi on opetettu jäsentämään sisällöt ihmisten käyttämien tekstilajien (genre) perusteella, kuten ne tekoälyalgoritmien opetusaineistossa ilmenevät. Tekoälyalgoritmi ainoastaan hakee todennäköisyysanalyysiin perustuen aineistosta merkkiä tai sanaa, joka todennäköisimmin kuuluu merkkijonon seuraavaksi osaksi. (Gherson 2023, 118.) Gherson muotoilee samalla esiin keskeisen eron inhimillisen ajattelun ja AI-palvelun tulosten välillä: Ihmisille tieto on aina kontekstoitua ja sillä on maailmassa viittauskohteita. Sen sijaan algoritmit rakentavat tuloksia opetusaineistossa ilmenevien todennäköisyyksien avulla.

Todennäköisyyksien tarkastelu nostaa keskiöön kysymyksen opetusaineistoista, joita algoritmien opettamiseen on käytetty. Esimerkiksi OpenAI on käyttänyt AI-palvelujensa (mm. ChatGPT-versiot, Dall-E) opetusaineistona koko internetistä löytyvää sisältöä kuten myös Googlen Gemini. Samalla internet-sisältöihin liittyvät vinoumat ovat siirtyneet osaksi algoritmien toimintaa. Tämä tarkoittaa vaikkapa erilaisten kansallisuuteen, etnisyyteen, sukupuoleen, ihonväriin, ikään, koulutukseen, tulotasoon, poliittiseen kantaan tai asuinpaikkakuntaan liittyvien aineistovinoutumien toistumista promptausten tuloksissa, kuten on laajalti havaittu tapahtuneen. Esimerkiksi Googlen Geminin kuvageneraattorin tuottamia vinoumia yritettiin ratkoa erillisellä korjausalgoritmilla, mutta yritys epäonnistui räikeästi ja palvelu piti sulkea eikä ongelmia ole vielä toukokuussa 2024 saatu korjattua (Wiggers 2024a). Samankaltaisia syytöksiä on esitetty myös Microsoft Copilotin Designer-työkalusta ja Microsoftin Bing-hakuohjelmasta (Leppälä 2024; Perrigo 2023; Vincent 2023). Opetusaineistojen vinoumiin on havahduttu ja ohjelmistojätit työskentelevät eri tavoin tilanteen ratkaisemiseksi (Collective Labs 2024; Fadelli 2024). Opetusaineistoja laajennetaan ja niitä ajantasaistetaan uusiin ohjelmistoversioihin samalla, kun algoritmeja kehitetään, mistä kertoo sovellusten tiheä päivitysrytmi.

Kenelle kuuluvat opetusaineistot ja tulokset?

Algoritmien opetusaineistojen osalta on törmätty kysymyksiin tekijänoikeuksista ja aineistojen luvattomasta AI-opetuskäytöstä. Varsinkin kuvataiteilijat ja valokuvaajat ovat hermostuneet, kun verkossa olevia tekijänoikeussuojattuja kuva-aineistoja on käytetty tekijöiden tietämättä ja ilman lupaa tekoälyalgoritmien opetusaineistoina, minkä seurauksena näiden aineistojen elementtejä on eri tavoin muunneltuina ilmaantunut kuva- ja videogeneraattoreiden antamiin tuloksiin. Reaktiona tilanteeseen Chicagon yliopiston tutkijat ovat kehittäneet Nightshade- ja Glaze -sovellukset, joiden avulla kuva-aineistojen oikeuksienhaltijat voivat suojata verkossa olevia aineistojaan. Esimerkiksi Nightshade ”myrkyttää” kuvatiedostot lisäkoodilla, joka on ihmiskatsojalle näkymätöntä mutta se saa algoritmin tulkitsemaan kuvan sisällön väärin. Asiaa käsitelleen artikkelin esimerkissä Mona Lisaa esittävä kuva näkyy ihmiselle normaalisti, mutta algoritmi generoi esiin kissahahmon. Vastaavasti Glaze manipuloi taiteellista tyyliä, jolloin tavoitellun visuaalisen tyylin sijaan käskyn

tuloksena palautuu jotain muuta. (Sung 2024.) Selvää kuitenkin on, että tällaiset ratkaisut ovat vain osittaisia ja todennäköisesti myös tilapäisiä teknologioiden kehittyessä.

Nightshade ja Glaze ovat reaktioita aineistojen luvattomaan käyttöön ja niiden tuottamat ”hallusinaatiot” tulee ymmärtää tässä kontekstissa. Sen sijaan AI-opetusaineistot kohtaavat myös toisenlaisia haasteita eli tarkoituksellisia yrityksiä vääristää, sotkea tai ”myrkyttää” näitä aineistoja ja siten haitata palvelujen antamien tulosten oikeellisuutta. Näitä tekniikoita on kehitetty jo roskapostin tiimoilta (”kuinka v.i.a.g.r.a. saadaan läpi roskapostisuotimista”), mutta nykytilanteessa kyse on laajamittaisesta ja järjestäytyneestä toiminnasta, jossa käytetään kehittyneitä hyökkäys- ja tunkeutumiskeinoja sekä myös keinoja näiltä hyökkäyksiltä suojautumiseen (ks. lisää Wikipedia 2024). Oma aihealueensa on opetusaineistoja korruptoiva netin disinformaatiosisältö, jota suodatusyrityksistä huolimatta päätyy aineistoihin. Tiivistetysti voidaan todeta, että algoritmien opetusaineistoihin liittyy vielä lukuisia ongelmia, joista vain osa on teknisiä.

Opetusaineistojen laadun parantamiseksi ja tulosvääristymien vähentämiseksi keskeiset AI-jätit ovat alkaneet tehdä sopimuksia kansainvälisten julkaisutalojen ja aineistoalustojen kanssa, jotta niiden hallitsemien oikeuksien alaisia aineistoja saadaan opetusaineistoiksi. Esimerkiksi OpenAI on sopinut sosiaalisen median Reddit-uutissivuston aineistojen käytöstä AI-sovellusmalliensa opettamiseen (Wiggers 2024b). Toisaalla Sony Music on havahtunut tekoälyllä tuotetun musiikin nopeaan yleistymiseen musiikin verkkojakelussa ja kirjelmöi teknologiayhtiöitä varoittaen käyttämästä Sonyn aineistoja AI-opetukseen ilman lupaa (Malik 2024). Myös keskustelualusta Slack on kohdannut syytöksiä keskusteluaineistojen luvattomasta opetuskäytöstä (Mehta & Lunden 2024) samoin kuin Meta ja Instagram-alusta kuvien osalta (Komonen 2024c). Tekijänoikeuksien tehokkaan suojaamisen kannalta tilanne on monella tavalla menetetty, koska verkossa olevia sisältöjä on jo käytetty – ja käytetään koko ajan – tekoälysovellusten opetusaineistoina ilman, että sitä pystytään teknisesti estämään. Muodostuvassa tilanteessa koko tekijänoikeusjärjestelmä ja tekijänoikeuksien käsite alkaa olla syvissä ongelmissa, minkä tässä yhteydessä voi vain todeta: kuinka tekijänoikeus tulisi ylipäätään ymmärtää silloin, kun AI-avusteista tai kokonaan sen generoimaa sisältöä käytetään ja julkaistaan tekijänoikeusteollisuudessa?

Hallusinoinnilta ja tulosvääristymiltä pyritään myös suojautumaan käyttämällä toista sovellusta yhden sovelluksen tulosten varmistamiseen ja tähän tarkastussykliin voidaan liittää muitakin sovelluksia, jolloin virheiden mahdollisuus pienenee. Varsinaisesti tällaisilla toimilla ei kuitenkaan ole vaikutusta siihen, miten me palveluiden käyttäjät voimme itse arvioida promptiemme tuottamia tuloksia. Ne jäävät koko ajan läpinäkymättömiksi emmekä aidosti tiedä, miten niihin on päädytty, vaikka pyytäisimme palvelua esittämään lähteet, joihin tiedot perustuvat. Yhä kehittyneemmin, näyttävämmin ja vakuuttavammin muotoillut tulokset synnyttävät helpommin ihailua kuin kritiikkiä, ja kaikki tuo tapahtuu *maagisen* helposti. Saamme valmiita tuloksia mutta emme osaa lopulta sanoa, miten niihin on päädytty. Jos tuloksiin liittyy vääristymiä eli AI-hallusinaatioita, sovellusten kehittyessä meidän on yhä vaikeampi tunnistaa niitä muun sisällön joukosta. Koulutusjärjestelmän ja varsinkin korkeakoulutuksen näkökulmasta tähän liittyy haasteita, ja haasteita tähän toki liittyy kaikkien muidenkin tekoälypalveluihin turvautuvien toimijoiden kannalta.

Uhkia ja mahdollisuuksia

Kun tunnistamme tekoälysovelluksen antamassa tuloksessa jotakin outoa, ajattelemme oletusarvoisesti, että tekoäly hallusinoi tai satuilee. Tämä havainto on erityisen helppo tehdä vaikkapa kuvageneraattorien tulosten kohdalla, jos ihmisellä on kädessään seitsemän sormea. Aina tuon outouden ei tarvitse olla satuilua tai vääristymää. Algoritmit voivat myös yhdistellä tietoaineistoja innovatiivisesti sellaisella meille uudella tavalla, joka on todellisuudessa mahdollinen. Tällöin algoritmiikka toimii luovan ajattelun tukena ja välineenä. Joissakin kielimalleihin perustuvissa sovelluksissa tekoälyn ”luovuutta” voidaan säätää komennon yhteydessä, jolloin ”luovuutta” lisäämällä saadaan algoritmit hallusinoimaan voimakkaammin ja tulokset voivat olla villejä. Joissakin näistä tuloksista voi kuitenkin olla ominaisuuksia, jotka auttavat ihmiskäyttäjää luovassa työskentelyssä. Selvää on, että vastuu tällaisten tulosten tulkinnasta on käyttäjällä – tässä suhteessa tekoälysovellukset ovat kuin ”savolainen puhe”.

Tekoälysovellusten käytön räjähdysmäinen yleistyminen on nostanut esiin aiheellisen huolen palveluiden aiheuttamista kustannuksista ja

ympäristövaikutuksista. Yksinkertaistenkin promptausten toteuttaminen vaatii runsaasti laskentatehoa, jolloin sähköä kuluu. Arvion mukaan tekoälypalveluiden käyttö nostaa lähivuosina fossiilisten polttoaineiden kulutusta maailmassa useita prosentteja samalla, kun hiilidioksidipäästöt kasvavat merkittävästi (Kangasniemi 2024) – vaikka tekoälysovellukset selviäisivätkin tietyistä tehtävistä ihmisiä energiatehokkaammin (Tomlinson et al. 2024). Mikäli palvelut kehittyvät nykyisellä vauhdilla, mikä käytännössä tarkoittaa koko ajan suurempia opetusaineistoja ja lisääntyvää laskentakapasiteettia, niin esitetyt ympäristövaikutukset voivat olla vielä selvästi esitettyä suurempia. Kehitys vaikeuttaa energiatuotannon globaalia vihreää siirtymää.

Kirjoittaja:

Jari Kupiainen, yliopettaja, Karelia-ammattikorkeakoulu

Lähteet:

Collective Labs 2024: ”Why Spiral.” https://www.collectivelabs.ai/why-spiral. 27.5.2024.

Edwards, Benj 2023: “Why ChatGPT and Bing Chat Are So Good at Making Things Up.” *ArsTechnica*. 6.4.2023. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/. 27.5.2024.

Emsley, Robin 2023: ”ChatGPT: These Are Not Hallucinations – They’re Fabrications and Falsifications.“ *Schizophrenia*. 9:1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10439949/. 27.5.2024.

Fadelli, Ingrid 2024: “A Method to Mitigate Hallucinations in Large Language Models.” TechXplore. 22.5.2024. https://techxplore.com/news/2024-05-method-mitigate-hallucinations-large-language.html. 27.5.2024.

Gherson, Ilana 2023: “Bullshit Genres: What to Watch for When Studying the New Actant ChatGPT and Its Siblings.” *Suomen Antropologi*. 47:3. 115–131. https://journal.fi/suomenantropologi/article/view/137824/85841. 27.5.2024.

Gold, Tyler 2023: “Indistinguishable from Magic: Magic Is a Fun and Pragmatic Lens for Understanding Generative AI — Cast a Spell and Conjure Content.“ *Gold’s Guide*. 15.1.2023. https://goldsguide.com/indistinguishable-from-magic/. 27.5.2024.

Huttunen, Matti 2018: ”Harha-aistimus (hallusinaatio).” *Lääkärikirja Duodecim*. https://www.terveyskirjasto.fi/dlk00371. 27.5.2024.

Kangasniemi Tuomas 2024: ”Järkyttävä ennuste: Fossiilisten polttoaineiden kulutus nousemassa rajusti 2025–30, kun tekoäly nielee valtavasti sähköä – Päästöjä jopa 900 000 000 tonnia lisää vuodessa.” *Tekniikka & talous*. 29.4.2024. https://www.tekniikkatalous.fi/uutiset/tt/5a5fb2aa-8677-4217-85c2-a75ac46829c7. 27.5.2024.

Keary, Tim 2024: “AI Hallucination.” *Techopedia*. 15.1.2024. https://www.techopedia.com/definition/ai-hallucination. 27.5.2024.

Komonen, Joona 2024a: “Onko tässä tekoälyn tulevaisuus?”. *Mikrobitti*. 15.5.2024. https://www.mikrobitti.fi/uutiset/mb/cbb249cd-124c-4429-80c4-327c60918b4a?ref=ampparit:7e84&_gl=1*9xcrzu*_ga*MTYyMjA2OTYxNS4xNzE1MzI3MTA4*_ga_3L539PMN3X*MTcxNTc3NTYxOS4xMy4wLjE3MTU3NzU2MTkuMC4wLjA. 27.5.2024.

Komonen, Joona 2024b: “Microsoft kehittää oman tekoälymallin – haastaa Geminin ja GPT-4:n.” *Tivi*. 10.5.2024. https://www.tivi.fi/uutiset/microsoft-kehittaa-oman-tekoalymallin-haastaa-geminin-ja-gpt-4n/0f008869-2ae0-49d3-ac68-12ce16b654c8. 27.5.2024.

Komonen, Joona 2024c: ”Pomo myönsi: Instagram-kuvia käytetään tekoälyn koulutukseen – näin voi suojata kuvansa.” Mikrobitti. 26.5.2024. https://www.mikrobitti.fi/uutiset/mb/cb00e7a9-4986-4cd1-b7c6-

420aabc5c09b?ref=ampparit:e219&_gl=1*19s2foh*_ga*MTYyMjA2OTYxNS4xNzE1MzI3MTA4*_ga_3L539PMN3X*MTcxNjcxODI4My4zOS4wLjE3MTY3MTgyODMuMC4wLjA. 27.5.2024.

Leffer, Lauren 2024: "AI Chatbots Will Never Stop Hallucinating." *Scientific American*. 5.4.2024. https://www.scientificamerican.com/article/chatbot-hallucinations-inevitable/. 27.5.2024.

Leppälä, Samuli 2024: "Microsoft's Artificial Intelligence Application Accused of Indecency." *Tivi*. 9.3.2024. https://www.tivi.fi/uutiset/microsofts-artificial-intelligence-application-accused-of-indecency/d60a3908-8a5c-4d1a-8b65-4f6d897380ba. 27.5.2024.

Lomas, Natasha 2024: "ChatGPT's 'Hallucination' Problem Hit with Another Privacy Complaint in EU." *TechCrunch*. 29.4.2024. https://techcrunch.com/2024/04/28/chatgpt-gdpr-complaint-noyb/. 27.5.2024.

Maleki, Negar, Balaji Badmanabhan & Kaushik Dutta 2024: "AI Hallucinations: A Misnomer Worth Clarifying." *arXiv.* 2401.06796v1. 9.1.2024. https://arxiv.org/html/2401.06796v1. 27.5.2024.

Malik, Aisha 2024: "Sony Music Warns Tech Companies over 'Unauthorized' Use of Its Content to Train AI." *TechCrunch*. 16.5.2024. https://techcrunch.com/2024/05/16/sony-music-warns-tech-companies-over-unauthorized-use-of-its-content-to-train-ai/. 27.5.2024.

Mehta, Ivan & Ingrid Lunden 2024: "Slack Under Attack Over Sneaky AI Training Policy." *TechCrunch*. 17.5.2024. https://techcrunch.com/2024/05/17/slack-under-attack-over-sneaky-ai-training-policy/. 27.5.2024.

Perrigo, Billy 2023: "The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter." *Time*. 17.2.2023. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/. 27.5.2024.

Puttonen, Mikko 2024: "Tietoisuuden tutkija: 'Elämme hallitussa hallusinaatiossa'." *Helsingin Sanomat*. 25.4.2024. B 7.

Pölönen, Harri 2024: “AI-biturientti: tekoäly yo-kokeessa.” *Skrolli*. 1/2014. 10–14.

Ruckenstein, Minna 2022: *The Feel of Algorithms*. Oakland, CA.: University of California Press.

Simonite, Tom 2018: “AI Has a Hallucination Problem That’s Proving Tough to Fix.” *Wired*. 9.3.2018. https://www.wired.com/story/ai-has-a-hallucination-problem-thats-proving-tough-to-fix/. 27.5.2024.

Sung, Morgan 2024: “Nightshade, the Tool that ‘Poisons’ Data, Gives Artists a Fighting Chance Against AI.” *TechCrunch*. 26.1.2024. https://techcrunch.com/2024/01/26/nightshade-the-tool-that-poisons-data-gives-artists-a-fighting-chance-against-ai/. 27.5.2024.

Tomlinson, B., R.W. Black, D.J. Patterson *et al.* 2024: “The Carbon Emissions of Writing and Illustrating Are Lower for AI than for Humans.” *Scientific Reports*. 14: 3732. https://doi.org/10.1038/s41598-024-54271-x. 27.5.2024.

Vincent, James 2023: “Microsoft’s Bing Is an Emotionally Manipulative Liar, and People Love It.” *The Verge*. 15.2.2023. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams. 27.5.2024.

Wasilewski, Eryk & Mirek Jablonski 2024: “Measuring the Perceived IQ of Multimodal Large Language Models Using Standardized IQ Tests.” *TexRxiv*. 13.5.2024. https://www.techrxiv.org/doi/full/10.36227/techrxiv.171560572.29045385/v1. 27.5.2024.

Wiggers, Kyle 2024a: “Google Still Hasn’t Fixed Gemini’s Biased Image Generator.” *TechCrunch*. 15.5.2024. https://techcrunch.com/2024/05/15/google-still-hasnt-fixed-geminis-biased-image-generator/. 27.5.2024.

Wiggers, Kyle 2024b: “OpenAI Inks Deal to Train AI on Reddit Data.” *TechCrunch*. 16.5.2024. https://techcrunch.com/2024/05/16/openai-inks-deal-to-train-ai-on-reddit-data/. 27.5.2024.

Wikipedia 2024: “Adversarial Machine Learning.” Wikipedia. 21.5.2024. https://en.wikipedia.org/wiki/Adversarial_machine_learning. 27.5.2024.