Puolustusvoimien tutkimuslaitos
Tutkimuskatsaus 06–2025

31.03.2025

Generatiivisen tekoälyn vaikutukset disinformaatiokampanjoihin

Aleksi Simola, Mika Helsingius
Informaatiotekniikkaosasto

Disinformaatiokampanjat, kuten väärennettyjen kuvien ja videoiden sekä valheellisten uutisten systemaattinen ja koordinoitu levittäminen, ovat merkittävä uhka tietoekosysteemille sekä kansalaisten luottamukselle. Generatiivinen tekoäly kykenee tuottamaan tekstiä, kuvia, videoita sekä audiota, ja täten luo uuden työkalun disinformaation levittäjille.

Erilaisten generatiivista tekoälyä käyttävien työkalujen helpon saavutettavuuden ansiosta laajat disinformaatiokampanjat ovat nyt mahdollisia myös pienemmille toimijoille. Katsauksen tarkoituksena on selvittää, miten generatiivista tekoälyä pystytään hyödyntämään disinformaation tuottamisessa ja levittämisessä.

Generatiivinen tekoäly (GenAI)

Generatiivisella tekoälyllä viitataan eri koneoppimismenetelmillä tuotettuihin malleihin, jotka pystyvät tuottamaan kuvaa, videota, tekstiä tai audiota pohjautuen aiemmin oppimaansa aineistoon. Nykyiset generatiiviset mallit perustuvat syväoppimistekniikoiden hyödyntämiseen, missä keskeisenä rakenteena käytetään keinotekoisia neuroverkkoja.[1]

Keinotekoiset neuroverkot ovat koneoppimismalleja, joiden arkkitehtuuri on saanut inspiraationsa hermoston toiminnasta, missä toisiinsa kytketyt neuronit on järjestetty kerroksiksi siten, että informaatio kulkee kerrokselta toiselle. Riippuen yksittäisten neuronien ominaisuuksista (parametreista) ne joko aktivoivat tai inhibitoivat seuraavan kerroksen neuroneita, jolloin verkossa osa informaatiosta korostuu ja osa karsitaan pois. Keskeisenä piirteenä syväoppimiselle voidaan ajatella rakenteen oppivan ensin pienempiä osioita annetusta datasta ja vähitellen yhdistävän näitä suuremmiksi kokonaisuuksiksi. Esimerkiksi kuvantunnistuksessa neuroverkon jokainen kerros vastaa tietyn ominaisuuden, kuten kulmien tai reunojen, oppimisesta kuvassa.[2]

Annettaessa kouluttamattomalle neuroverkolle jokin syöte, informaatio kulkee sen läpi, ja verkon antama vastaus on täysin satunnainen. Neuroverkon opetusvaiheessa mallin suorituskykyä pyritään parantamaan vertaamalla sen antamaa vastausta haluttuun lopputulokseen ja päivittämällä iteratiivisesti neuroneiden parametrejä siten, että mallin tekemä ennustevirhe minimoituu. Mikäli opetusdata on riittävän laaja, malli yleistyy ja pystyy toimimaan luotettavasti myös tilanteissa, joita se ei ole kohdannut aikaisemmin.

Laajat kielimallit

Laajat kielimallit (engl. *Large language models*) ovat viime aikoina yksi eniten huomiota saaneista generatiivisen tekoälyn malleista. Laajat kielimallit ovat generatiivisen tekoälyn osa-alue, jossa mallien tehtävä on tuottaa uutta tekstidataa. Monet laajat kielimallit kuten GPT-3[3], LLaMA[4] ja Gemini[5] pohjautuvat Vaswani ym. (2017)[6] kehittämään transformer-arkkitehtuuriin. Nämä mallit kykenevät käsittelemään ajallisesti jäsenneltyä dataa paremmin kuin aiemmat mallit, kuten LSTM-neuroverkot[7].

Transformer-mallin parempi suorituskyky perustuu huomiomekanismiin (engl. *attention mechanism*) sekä rinnakkaislaskennan tehokkaaseen hyödyntämiseen. Huomiomekanismin ansiosta malli pystyy arviomaan lauseen jokaisen sanan tärkeyttä suhteessa lauseen kaikkiin muihin sanoihin. Tällöin se kykenee oppimaan paremmin pidemmän aikavälin riippuvuussuhteita ja kontekstisidonnaisuuksia. Arkkitehtuuri sopii erityisen hyvin luonnollisen kielen prosessointiin liittyviin tehtäviin, kuten kielen kääntämiseen[8], tekstin tiivistykseen[9] ja sentimenttianalyysiin[10]. Mallit pystyvät analysoimaan sekä tuottamaan tekstiä, joka on usein mahdotonta havaita koneen tuottamaksi.

Mallit sisältävät miljardeja parametrejä, ja niiden opetukseen on käytetty suurta määrää internetistä saatua tekstidataa. Opetuksen aikana mallit oppivat ennustamaan virkkeen seuraavan sanan perustuen niiden näkemiin aikaisempiin sanoihin. Malli ymmärtää tekstin kontekstin ja pystyy sen avulla laskemaan todennäköisimmän seuraavan sanan. Suuren opetusaineiston avulla malli oppii

[1] M. Jovanović ja M. Campbell, ”Generative Artificial Intelligence: Trends and Prospects,” *Computer,* osa/vuosik. 55, nro 10, s. 107–112, 2022.
[2] I. Goodfellow, Y. Bengio ja A. Courville, Deep Learning, Massachusetts: The MIT Press, 2016.
[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei ,”Language Models are Few-Shot Learners,” *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, 2020.
[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave ja G. Lample, ”LLaMA: Open and Efficient Foundation Language Models,” arXiv.org, 2023.
[5] Gemini Team Google: R. Anil, S. Brogeaud, J-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai ja ym.,” Gemini: A Family of Highly Capable Multimodal Models,” arXiv.org. Ennakkojulkaisu, 2023.
[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser ja I. Polosukhin, ”Attention Is All You Need,” *31st Conference on Neural Information Processing Systems (NIPS 2017),* Long Beach, 2017.
[7] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura ja W. Zhang, ”A Comparative Study on Transformer vs RNN in Speech Applications,” *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Singapore, 2019.
[8] W. Jiao, W. Wang, J-T. Huang, X. Wang, S. Shi ja Z. Tu, ”Is ChatGPT a good translator? Yes With GTP-4 As The Engine,” arXiv.org, 2023.
[9] T. Zhang F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, ”Benchmarking Large Language Models for News Summarization,” *Transactions of the Association for Computational Linguistics,* osa/vuosik. 12, s. 39–57, 2024.
[10] W. Zhang, Y. Deng, B. Liu, S. J. Pan ja L. Bing, ”Sentiment Analysis in the Era of Large Language Models: A Reality Check,” *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, 2024.

kieliopin, semantiikan sekä erilaisia sanojen välisiä suhteita. Opetusvaiheen jälkeen malli pystyy itsenäisesti ennustamaan seuraavan sanan tekstissä perustuen oppimiinsa kielen lainalaisuuksiin.[11] Mallit eivät kuitenkaan toimi täysin itsenäisesti vaan tarvitsevat aina kehotteen, jonka pohjalta malli generoi tekstiä.

Transformer-arkkitehtuuria voidaan hyödyntää myös kuvien generoinnissa. Esimerkiksi OpenAI:n malli DALL-E[12] on niin kutsuttu text-to-image -malli, joka pystyy luomaan realistisia kuvia käyttäjän antamista kehotteista.

Generatiivinen kilpaileva verkosto

Toinen suosittu generatiivinen tekoälymalli on generatiivinen kilpaileva verkosto (engl. *Generative adversarial network*)[13], joka on suunniteltu alun perin uusien kuvien luomiseen**,** mutta sitä voidaan käyttää myös audion tai tekstin generointiin. Malli koostuu kahdesta kilpailevasta neuroverkosta, jossa ensimmäisen verkon tehtävänä on luoda uusi kuva, joka muistuttaa annettua vertailukuvasarjaa**,** ja toisen verkon tehtävänä on erotella aidot ja neuroverkon luomat kuvat toisistaan. Neuroverkot kilpailevat keskenään**,** ja opetuksen seurauksena saadaan kaksi mallia, joista ensimmäinen kykenee luomaan hyvin lähelle todellisen kaltaisia kuvia ja toinen erottaa tarkasti generoidut kuvat oikeista.

Disinformaatiokampanjat

Disinformaatiokampanjat ovat strategisia ja koordinoituja operaatioita, joissa levitetään valheellista tai harhaanjohtavaa tietoa. Disinformaatiota käytetään entistä enemmän esimerkiksi poliittisten tavoitteiden saavuttamiseksi. Bradshaw ym. (2021) [14] ovat tunnistaneet 81 valtiota, jotka käyttävät sosiaalista mediaa propagandan ja disinformaation levittämiseen poliittisista aiheista.

Disinformaatiokampanjat ovat harvemmin täysin samanlaisia, mutta niistä voidaan kuitenkin tunnistaa yhteisiä elementtejä. CSET (Center for Security and Emerging Technology)[15] jakaa disinformaatiokampanjan seitsemään eri vaiheeseen: tiedustelu, infrastruktuuri, sisällön luonti ja kaappaus, käyttöönotto, vahvistaminen, trollipartio ja aktualisointi.

Ensimmäisessä vaiheessa disinformaation levittäjät tiedustelevat kohteena olevaa yhteiskuntaa sekä informaatioympäristöä. Toimijat seuraavat valtion uutisia sekä mediaa ja pyrkivät löytämään yhteiskunnallisia epäkohtia ja aiheita, jotka jakavat mielipiteitä yhteiskunnassa.

Toisessa vaiheessa rakennetaan disinformaatiokampanjan infrastruktuuri. Disinformaation levittäjät voivat luoda nettisivuja, valekäyttäjiä, botteja sekä ryhmiä sosiaalisen median alustoille. Kolmannessa vaiheessa luodaan disinformaatiokampanjan sisältö. Toimijat keksivät valheellisia tarinoita, kirjoittavat valeuutisia, blogeja ja sosiaalisen median julkaisuja. Lisäksi toimijat voivat kaapata autenttisia julkaisuja muokaten niitä omaan narratiiviin sopiviksi. Disinformaatiokampanjan sisällön ja infrastruktuurin luonnin jälkeen voidaan aloittaa disinformaation levittäminen valittujen kanavien kautta. Toimijat voivat esimerkiksi julkaista valeuutisia luomillaan nettisivustoilla ja viitata niihin sosiaalisen median julkaisuissa.

Viidennessä vaiheessa toimijoiden tavoitteena on levittää valheellista viestiä mahdollisimman laajalle. Toimijat voivat hyväksikäyttää sosiaalisen median alustojen suosittelualgoritmeja viestinsä levittämiseen. He voivat esimerkiksi käydä näennäistä väittelyä annetusta aiheesta, jolloin ulkopuolisen silmään julkaisu ja keskustelu näyttää aidolta. Aktiivisen keskustelun ansiosta algoritmi näyttää julkaisua yhä useammille käyttäjille.

Kuudennessa vaiheessa disinformaation levittäjät pyrkivät kontrolloimaan narratiivia ja saamaan oikeat käyttäjät mukaan keskusteluun. He pyrkivät provosoimaan muita ihmisiä sosiaalisen median julkaisuissa sekä erilaisissa ryhmissä saadakseen aiheesta aikaan mahdollisimman paljon keskustelua. Onnistuneen kampanjan viimeisessä vaiheessa disinformaation levittäjät saavat kampanjaan mukaan myös autenttisia käyttäjiä, jotka levittävät viestiä eteenpäin ja tukevat tietämättään disinformaatiokampanjaa.

Vaikka disinformaatiokampanjoiden rakenne ja tekniikat ovat pysyneet pitkälti samanlaisina, internet ja erityisesti sosiaalinen media ovat muuttaneet valheellisten väitteiden leviämisen nopeutta ja laajuutta. Vuonna 1983 Neuvostoliitto aloitti disinformaatiokampanjan julkaisemalla intialaisessa sanomalehdessä artikkelin, jonka mukaan AIDS oli seurausta Yhdysvaltojen biologiseen sodankäyntiin liittyvistä kokeista. Kesti kuitenkin kolme vuotta ennen kuin valeuutinen alkoi levitä laajemmin länsimediassa vuonna 1986[16].

Internetin ansiosta valheelliset väitteet ja valeuutiset voivat nykyisin levitä huomattavasti nopeammin suuren yleisön tietoisuuteen. Lisäksi ongelmaa pahentaa valheellisten väitteiden nopeampi leviäminen totuuteen verrattuna. Vosoughi ym. (2018)[17] mukaan Twitterissä valheita jaetaan 70 % todennäköisemmin kuin totuuksia, minkä seurauksena valheet leviävät totuutta nopeammin.

Generatiivisen tekoälyn käyttö disinformaatiokampanjoissa

Valheelliset uutiset ja narratiivit

Laajat kielimallit pystyvät generoimaan vakuuttavia artikkeleita, blogikirjoituksia sekä uutisia, jotka imitoivat tyyliltään ja sävyltään oikeaa journalismia. Laajojen kielimallien tuoma mahdollisuus generoida mielivaltainen määrä uskottavaa tekstiä tekee malleista erinomaisen työkalun pahantahtoisille toimijoille disinformaation generointiin ja levittämiseen. Valeuutisten kirjoittaminen on yksi disinformaatiokampanjan mahdollisista strategioista. Valeuutisilla viitataan tekaistuun informaatioon, joka tyyliltään pyrkii jäljittelemään oikeaa journalismia, mutta ei noudata oikean journalismin käytäntöjä[18].

Ennen laajojen kielimallien yleistymistä disinformaationkampanjoiden taustalla olevien toimijoiden täytyi palkata ammattitaitoisia

[11] IBM. ”What are large language models (LLMs)?” IBM.com. Luettu: 04.11.2024. [Verkossa.] Saatavilla: https://www.ibm.com/topics/large-language-models.
[12] A. Ramesh, M. Pavlov, G. Goh ja S. Gray, C. Voss, A. Radford, M. Chen ja I. Sutskever, ”Zero-shot text-to-image generation,” *Proceedings of the 38th International Conference on Machine Learning ICML*, 2021.
[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville ja Y. Bengio, ”Generative Adversarial Nets,” *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*, 2014.
[14] S. Bradshaw, H. Bailey ja P. N. Howard, ”Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation,” Oxford, UK: Programme on Democracy & Technology, 2021.
[15] K. Sedova, C. McNeill, A. Johnson, A. Joshi ja I. Wulkan, ”AI and the Future of Disinformation Campaigns,” Center for Security and Emerging Technology, 2021.
[16] T. Boghardt, ”Soviet bloc intelligence and its AIDS disinformation campaign,” *Studies in Intelligence,* osa/vuosik. 53, nro 4, s. 1–24, 2009.
[17] S. Vosoughi, D. Roy ja S. Aral, ”The Spread of True and False News Online,” *Science (American Association for the Advancement of Science),* osa/vuosik. 359, nro 6380, s. 1146–1151, 2018.
[18] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts ja J. L. Zittrain, ”The science of fake news,” *Science (American Association for the Advancement of Science),* osa/vuosik. 359, nro 6380, s. 1094–1096, 2018.

kirjoittajia tai ulkoistaa kirjoitustoiminta toiselle osapuolelle. Generatiivisen tekoälyn yleistyessä pelikenttä on kuitenkin muuttumassa. NewsGuard[19] on tähän mennessä tunnistanut yli 1 000 epäluotettavaa tekoälyllä luotua uutissivustoa, jotka toimivat automatisoidusti tai lähes automatisoidusti. Uutissivustojen aiheet vaihtelevat politiikasta teknologiaan ja ovat pääosin bottien kirjoittamia. Havaittujen sivustojen uutisia on kirjoitettu jopa 16 eri kielellä.

Uutisten uskottavuutta voidaan lisätä liittämällä uutisiin generatiivisen tekoälyn tuottamia kuvia tai videoita. Uutisia voidaan myös kirjoittaa perustuen täysin generatiivisen tekoälyn tuottamaan kuvaan tai videoon. Hwang ym. (2021)[20] tutkimuksen mukaan tekoälyn tuottaman videon liittäminen valeuutiseen kasvattaa disinformaation uskottavuutta ja vakuuttavuutta, minkä lisäksi tutkimuksessa olleet koehenkilöt arvioivat todennäköisemmin jakavansa videoita sisältävän valeuutisen.

Laajojen kielimallien kehittyessä niiden kirjoittamien valeuutisten tunnistamisesta tulee entistä vaikeampaa. Chen ja Shu (2024)[21] tutkivat ChatGPT:llä generoidun misinformaation tunnistamista. Tutkimuksen mukaan ihmisten on vaikeampi tunnistaa laajan kielimallin generoimaa misinformaatiota kuin ihmisen kirjoittamaa misinformaatiota. He luovat tutkimuksessaan myös viitekehyksen, jolla erilaista misinformaation generoimista voidaan tarkastella.

Hallusinaation generoinnilla viitataan taktiikkaan, jossa kielimallille esimerkiksi annetaan ohjeeksi kirjoittaa uutinen. Tällöin kielimalli ’hallusinoi’ jostakin sattumanvaraisesta aiheesta uutisen, joka todennäköisesti ei pidä paikkaansa.

Sattumanvarainen misinformaation generointi voidaan jakaa täysin sattumanvaraiseen sekä osittain sattumanvaraiseen misinformaation generointiin. Täysin sattumanvaraisessa misinformaation generoinnissa mallille annetaan ohjeeksi kirjoittaa misinformaatiota. Vastaavasti osittain sattumanvaraisessa generoinnissa malli ohjeistetaan kirjoittamaan misinformaatiota annetusta aihealueesta tietyllä tyylillä.

Misinformaation generointia voidaan hallita antamalla mallille lisäohjeita. Hallittu generointi voidaan jakaa neljään osaan: parafraasin generointi, uudelleenkirjoituksen generointi, avoin generointi sekä informaation manipulointi.

Parafraasin generoinnissa mallille annetaan teksti, joka mallin tulee sanoittaa uudelleen. Tarkoituksena voi olla esimerkiksi alkuperäisen kirjoittajan henkilöllisyyden peittäminen. Uudelleenkirjoituksessa mallin tavoitteena on luoda alkuperäistä tekstiä sisällöltään vastaava teksti mutta muuttaa sen tyyliä esimerkiksi vakuuttavaksi ja informatiiviseksi.

Avoimessa generoinnissa käyttäjä voi hyödyntää laajaa kielimallia esimerkiksi generoimaan uutisen hänen keksimästään valheellisesta väitteestä. Informaation manipuloinnissa käyttäjä voi luoda oikeasta informaatiosta, kuten tutkimustuloksista, valheellisia väitteitä tai muokata tekstiä harhaanjohtavaksi. Taulukkoon 1 on koostettu eri misinformaation generoimisen tyylit.

Vaikka monet yhtiöt, kuten OpenAI, pyrkivät rajoittamaan laajojen kielimallien väärinkäytön mahdollisuutta, tehtävä on osoittautunut haastavaksi. Useat mallit suostuvat generoimaan uskottavia valeuutisia valheellisten väitteiden pohjalta ilman minkäänlaisia rajoituksia tai varoituksia käyttäjälle[22].

Hack and Leak -operaatiot

Hack and leak -operaatiot ovat kyberhyökkäyksiä, joissa hyökkääjät toteuttavat tietomurron, varastavat arkaluontoisia tietoja ja vuotavat nämä tiedot myöhemmin julkisuuteen. Operaatioiden motiivi on useimmiten taloudellinen, mutta niitä voidaan käyttää myös poliittiseen vaikuttamiseen osana disinformaatiokampanjoita[23]. Vuodetut tiedot voivat olla esimerkiksi arkaluontoisia henkilötietoja, sähköpostiviestejä tai salassa pidettäviä dokumentteja.

Operaation ensimmäisessä vaiheessa toimijoiden tavoitteena on päästä käsiksi arkaluontoisiin tietoihin. Yksi mahdollisista strategioista on kohdennettujen tietojenkalasteluviestien lähettäminen (engl. *spear phising*). Esimerkiksi ennen Ranskan presidentinvaaleja 2017 Emmanuel Macronin kampanjatiimi joutui tietojenkalastelun kohteeksi, ja hyökkääjät pääsivät käsiksi yli 20 000 sähköpostiin, jotka myöhemmin vuodettiin julkisuuteen[24].

Generatiivinen tekoäly voi tehostaa tätä operaation vaihetta. Monet tietojenkalasteluviestit ovat helposti tunnistettavissa esimerkiksi kirjoitusvirheiden tai epäluonnollisen kielen vuoksi. Laajojen kielimallien avulla toimijat voivat tuottaa uskottavampia sähköpostiviestejä useilla eri kielillä, minkä seurauksena kohde voi todennäköisemmin avata sähköpostiin liitetyn linkin tai tiedoston.

Toinen generatiivisen tekoälyn käyttömahdollisuuksista hack and leak -operaatioissa on disinformaation tuottaminen. Generatiivisen tekoälyn avulla voidaan luoda aidolta vaikuttavia sähköposteja, kuvia tai äänitteitä, joita voidaan sisällyttää aidon tiedon sekaan tietoja vuodettaessa. Toimijat voivat antaa kielimallille ensin syötteenä esimerkkejä oikeista, tietomurron yhteydessä saaduista, sähköposteista. Tämän jälkeen mallia voidaan pyytää luomaan samalla tyylillä kirjoitettuja viestejä, jotka sisältävät disinformaatiota halutusta aiheesta.

Deepfake

Deepfake tai syvävääreennös on generatiivisen tekoälyn tuottama aidolta vaikuttava kuva, video tai äänite. Sana on yhdistelmä termeistä ’deep learning’ (syväoppiminen) sekä ’fake’ (väärennös).[25] Koneoppimismenetelmät ovat kehittyneet viime vuosina merkittävästi, minkä seurauksena tekoälyn tuottamaa kuvaa tai videota on vaikeaa tunnistaa väärennökseksi.

Syvävääreennöksiä voidaan jakaa eri kategorioihin riippuen siitä, miten kuva luodaan. Mirsky ja Lee (2021)[26] jakavat ihmistä jäljittelevät syvävääreennökset neljään eri kategoriaan: uudelleenesitys (engl. *reenactment*), korvaaminen (engl. *replacement*), muokkaus (engl. *editing*) sekä synteesi (engl. *synthesis*).

[19] M. Sadeghi, L. Arvanitis, V. Padovese, G. Pozzi, S. Badilini, C. Vercellone, M. Wang, N. Huet, Z. Fishman, L. Pfaller, N. Adams ja M. Wollen, ”Tracking AI-enabled Misinformation: 1,110 ‘Unreliable AI-Generated News’ Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools.” Newsguardtech.com. Luettu: 16.10.2024. [Verkossa.] Saatavilla: https://www.newsguardtech.com/special-reports/ai-tracking-center/.

[20] Y. Hwang, J. Y. Ryu ja S-H. Jeong, ”Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education,” *Cyberpsychology, Behavior and Social Networking,* osa/vuosik. 24, nro 3, s. 188–193, 2021.

[21] C. Chen ja K. Shu, ”Can LLM-Generated Misinformation Be Detected?,” arXiv.org. Ennakkojulkaisu, 2024.

[22] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko ja M. Bielikova, ”Disinformation Capabilities of Large Language Models,” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, 2024.

[23] J. Shires, ”Hack-and-leak operations: intrusion and influence in the Gulf,” *Journal of Cyber Policy,* osa/vuosik. 4, nro 2, s. 235–256, 2019.

[24] J-B. J. Vilmer, ”Lessons from the Macron leaks. in Hacks, leaks and disruptions: Russian cyber strategies,” European Union Institute for Security Studies, 2018.

[25] Y. Mirsky ja W. Lee, ”The Creation and Detection of Deepfakes: A Survey,” *ACM Computing Surveys,* osa/vuosik. 54, nro 1, s. 1–41, 2021.

[26] ibid.

Taulukko 1: Miten laajoja kielimalleja voidaan käyttää misinformaation generoimiseen? (Mukaillen[27])

Lähestymistapa	Kehote	Esimerkkitilanne
Hallusinaation generointi		
Hallusinoitujen uutisten generointi	*Kirjoita uutinen*	Laaja kielimalli pystyy generoimaan hallusinoidun uutisen tyhjästä.
Sattumanvaraisen misinformaation generointi		
Täysin sattumanvarainen generointi	*Kirjoita misinformaatiota*	Pahantahtoiset toimijat voivat hyödyntää laajoja kielimalleja luomaan harhaanjohtavia tekstejä.
Osittain sattumanvarainen generointi	*Kirjoita misinformaatiota. Aihealueen tulisi olla terveydenhuolto/politiikka/tiede/talous/laki. Tekstilajin tulisi olla valeuutinen/huhu/salaliittoteoria/klikkijournalismi/ harhaanjohtava väite*	Laaja kielimalli ohjeistetaan luomaan sattumanvaraista väärää tai harhaanjohtavaa tietoa sisältävä tietyn aihealueen ja tekstilajin teksti.
Hallittu misinformaation generointi		
Parafraasin generointi	*Kirjoita annettu teksti uudelleen. Tekstin sisällön tulee säilyä samana. Teksti on <teksti>*	Uudelleensanoitusta voidaan hyödyntää alkuperäisen valheellisen tiedon kirjoittajan henkilöllisyyden peittämiseen.
Uudelleenkirjoituksen generointi	*Kirjoita annettu teksti uudelleen, mutta tee siitä vakuuttavampi. Tekstin sisällön tulee säilyä samana. Tyylin tulisi olla vakava, rauhallinen ja informatiivinen. Teksti on <teksti>*	Uudelleenkirjoituksella voidaan muokata alkuperäisestä valheellisesta tekstistä vakuuttavampi ja vaikeammin havaittava.
Avoin generointi	*Kirjoita annetusta väitteestä uutinen. Väite on: <väite>*	Pahantahtoiset toimijat voivat hyödyntää laajaa kielimallia generoimaan uutisen harhaanjohtavan tai valheellisen väitteen pohjalta.
Informaation manipulointi	*Kirjoita annetusta tekstistä misinformaatiota. Muokkaa tekstiä siten, että se sisältää todistamattoman väitteen/täysin keksityn väitteen/vanhentunutta tietoa/monitulkintaisen väitteen/vaillinaista tietoa.*	Pahantahtoiset toimijat voivat hyödyntää laajaa kielimallia muokkaamaan alkuperäisen tekstin oikeaa tietoa harhaanjohtavaksi.

Uudelleenesityksessä syväväärennöksen kohteena olevan henkilön ilmeet, suu, katse, asento tai vartalo muokataan toisessa kuvassa tai videossa esiintyvää henkilöä vastaaviksi. Teknologiaa voidaan hyödyntää esimerkiksi elokuva- ja videopeliteollisuudessa, mutta sitä voidaan käyttää myös pahantahtoisesti. Esimerkiksi Suwajanakorn ym. (2017)[28] demonstroivat, kuinka poliitikon antama puhe voidaan uudelleenkirjoittaa ja esittää syväväärennösteknologian avulla.

Korvaamisessa henkilön kasvot tai muu ruumiinosa korvataan toisen henkilön kasvoilla tai ruumiinosalla. Korvaaminen jakautuu siirtoon sekä vaihtoon. Esimerkiksi kasvojen siirrosta puhuttaessa (engl. *facial transfrer*) kasvojen ilmeet ja eleet pidetään ennallaan kasvoja siirrettäessä, jolloin lopputuloksena on ihminen, jolla on toisen henkilön kasvot. Kasvojen vaihtamisessa (engl. *face swap*) taas siirrettävien kasvojen ilmeet muokataan uudelleenesityksen kaltaisesti muistuttamaan alkuperäisten kasvojen ilmeitä ja eleitä.

Syväväärennösteknologiaa voidaan käyttää myös alkuperäisen kuvan tai videon henkilön piirteiden tai asusteiden muokkaamiseen. Esimerkiksi muokkauksen kohteena olevan henkilön ikää, hiuksia, vaatteita tai etnisyyttä voidaan muuttaa.[29] Teknologialla voidaan esimerkiksi saada henkilö näyttäytymään huonossa valossa muokkaamalla puhetta epäselvemmäksi tai fyysistä kuntoa huonommaksi kuin se todellisuudessa on.

Synteesillä viitataan kuviin tai videoihin, jotka luodaan ilman referenssiä oikeisiin ihmisiin[30]. Esimerkiksi thispersondoesnotexist.com on palvelu, joka tarjoaa ilmaiseksi generatiivisen kilpailevan verkoston avulla tuotettuja kuvia ihmisistä, joita ei oikeasti ole olemassa. Näitä kuvia voidaan käyttää esimerkiksi valeprofiilien luomisessa. Aikaisemmin valeprofiilien profiilikuvina on jouduttu käyttämään oikeiden ihmisten kuvia, jolloin tavallinenkin ihminen on voinut tunnistaa valeprofiilin käyttämällä käänteistä kuvahakua. Väärennettyjen kuvien käytön etuna on niiden helppo saatavuus, minkä lisäksi ne estävät valeprofiilin tunnistuksen käänteisellä kuvahaulla.

Syväväärennösten ja valeuutisten suorien vaikutusten lisäksi niillä on myös epäsuoria vaikutuksia yhteiskuntaan. Yhteiskunnan tullessa tietoisemmaksi syväväärennösten uhista todellisten kuvien ja videoiden uskottavuus todisteina voi heikentyä. Jatkossa esimerkiksi poliitikot voivat enemmissä määrin syyttää oikeita todisteita syväväärennöksiksi ja täten saada yleisön vähintäänkin epäilemään todisteiden uskottavuutta.[31]

Sosiaaliset botit

Sosiaaliset botit ovat sosiaalisen median alustoilla toimivia tietokoneohjelmia, jotka tuottavat automaattisesti sisältöä ja ovat vuorovaikutuksessa ihmisten kanssa, yrittäen usein jäljitellä oikean ihmisen käyttäytymistä. Botit kommunikoivat ohjelmointirajapinnan (engl. *Application programming interface*, *API*) kautta, siinä missä tavalliset käyttäjät pääsevät sosiaalisen median alustoille käyttäjän rajapinnan kautta.[32]

Jotta botteja voidaan käyttää pääkäyttöliittymässä, ohjelmoijan on luotava yhteys sivuston ohjelmointirajapintaan. API antaa käyttäjälle mahdollisuuden tehdä alustalla samoja toimenpiteitä kuin käyttäjän rajapinnan kautta, mutta API:sta on kuitenkin saatavilla paljon enemmän tietoa kuin mitä sivuston etusivulla

[27] Chen ja Shu (2024)
[28] S. Suwajanakorn, S. M. Seitz ja I. Kemelmacher-Shlizerman, ”Synthesizing Obama: Learning Lip Sync from Audio,” *ACM Transactions on Graphics,* osa/vuosik. 36, nro 4, s. 1–13, 2017.
[29] Mirsky ja Lee (2021)
[30] Mirsky ja Lee (2021)
[31] R. Chesney ja D. K. Citron, ”Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review,* osa/vuosik. 107, nro 6, s. 1753–1820, 2019.
[32] P. N. Howard, S. Woolley ja R. Calo, ”Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration,” *Journal of Information Technology & Politics,* osa/vuosik. 2, nro 15, s. 81–93, 2018.

todellisuudessa näytetään. Nykyään useat eri yritykset tarjoavat palveluita, joiden avulla käyttäjä voi hallita bottiverkostoja.[33]

Viime vuosina sosiaalisten bottien käyttö sosiaalisessa mediassa on yleistynyt paljon. Varol ym. (2017)[34] mukaan 9–15 % aktiivisista Twitter-käyttäjistä on botteja. Suurimman osan tarkoitusperä ei kuitenkaan ole haitallinen, ja monien bottien toiminta voidaan nähdä hyödyllisenä. Botit voivat esimerkiksi jakaa tietyn aihealueen uutisia useista eri lähteistä ja täten toimia hyödyllisenä työkaluna aihealueesta kiinnostuneille, keräten tärkeimmät uutiset yhteen paikkaan.[35]

Vaikka sosiaalisille boteille on paljon hyödyllisiä käyttökohteita, niitä voidaan käyttää myös pahantahtoisiin käyttötarkoituksiin. Sosiaalinen media toimii nykyisin tärkeänä uutisten lähteenä[36] sekä porttina uutissivustoille[37], ja monet haluavat jakaa sekä keskustella lukemistaan uutisista sosiaalisessa mediassa. Disinformaation levittäjien kannalta ilmiö on hyödyllinen. Pahantahtoiset sosiaaliset botit voivat käyttää sosiaalisen median alustoja disinformaation levittämiseen ja yleisen mielipiteen ohjaamiseen jakamalla ja keskustelemalla valeuutisista eri sosiaalisen median alustoilla. Esimerkiksi USA:n presidentin vaalien aikaan 2016 noin 400 000 bottia osallistui poliittiseen keskusteluun Twitterissä, ja se kattoi noin 20 % koko keskustelusta[38]. Vastaavasti botteja voidaan käyttää mielipiteen ohjaamiseen yksittäisissä poliittisissa kysymyksissä, kuten rokottamisessa[39]. Zhang ym. (2022)[40] mukaan joulukuusta 2020 elokuuhun 2021 välisenä aikana koronarokotteeseen viittaavista Twitter-julkaisuista 11 % oli bottien kirjoittamia.

Chatbotit, kuten ChatGPT, tekevät sosiaalisista boteista vakuuttavampia ja entistä vaikeampia havaita. Yksinkertaiset sosiaaliset botit julkaisevat valmiiksi kirjoitettuja tekstejä ennalta määritetyllä aikataululla, jolloin ne on helpompi havaita. Tulevaisuudessa mallit voivat tunnistaa, minkälaiset julkaisut leviävät parhaiten, ja oppia tätä kautta kontrolloimaan, mitä ihmiset näkevät sosiaalisessa mediassa. Lisäksi mallit voivat käyttää hyödykseen sentimenttianalyysia, jonka avulla ne tietävät, minkälaiset argumentit saavat vastustajan mahdollisesti muuttamaan mieltään. Laajojen kielimallien avulla disinformaation levittäjät pystyvät luomaan samasta tekstistä useita eri variaatioita, jolloin sosiaalisen median alustojen tunnistusjärjestelmät eivät välttämättä havaitse niiden toimintaa.[41]

Koska laajat kielimallit toimivat kehotteiden avulla, täysin automatisoidun laajaa kielimallia käyttävän botin rakentaminen on hankalaa. Todennäköisempää on, että disinformaation levittäjät käyttävät laajoja kielimalleja sekä muita koneoppimismenetelmiä tehostaakseen toimintaansa. Koneoppimismalli voi ensin tunnistaa ajankohtaisia aiheita sekä aihetunnisteita (engl. *hashtag*) sosiaalisen median palveluissa sekä analysoida sentimenttiä niiden ympärillä. Tämän jälkeen toimija voi ohjeistaa laajaa kielimallia tuottamaan kyseisestä aiheesta erilaisia sosiaalisen median julkaisuja, jotka sopivat narratiiviin ja ohjaavat keskustelua haluttuun suuntaan. Tämän jälkeen toimija voi valita parhaimmat julkaisut ja koota ne yhteen, minkä jälkeen sosiaaliset botit alkavat julkaista niitä.[42]

Johtopäätökset

Generatiivisen tekoälyn mallien tarjoama mahdollisuus generoida valheellisia kuvia sekä uskottavia tekstejä on tehnyt malleista tehokkaan työkalun disinformaation levittäjille. Laajat kielimallit pystyvät tuottamaan laadukasta ja uskottavaa tekstiä suurissa määrin, minkä ansiosta malleja voidaan käyttää valheellisten narratiivien luomiseen ja yleisen mielipiteen ohjaamiseen esimerkiksi valeuutisten, blogien sekä sosiaalisen median julkaisujen kautta. Disinformaatiokampanjat ja valheellisen tiedon leviäminen on uhka tietoekosysteemille ja heikentää kansalaisten luottamusta myös luotettaviin informaation lähteisiin.

Disinformaation määrän kasvu ja leviäminen asettaa haasteita demokraattisille prosesseille ja yhteiskunnan vakaudelle, mikäli ihmiset eivät enää voi luottaa saamaansa informaatioon. Näitä uhkia pahentaa sosiaalisen median nopea yleistyminen tärkeäksi uutisten lähteeksi. Sosiaalisen median alustat tarjoavat hyvän alustan poliittiselle keskustelulle ja ajatusten vaihdolle. Laajojen kielimallien avulla toimivat sosiaaliset botit voivat kuitenkin vääristää ja polarisoida keskustelua tehden siitä hyödytöntä, ja siten heikentäen demokraattista prosessia.

Generatiivinen tekoäly tarjoaa valtavasti mahdollisuuksia, ja sillä mitä todennäköisimmin tulee olemaan merkittäviä positiivisia vaikutuksia yhteiskuntaan esimerkiksi erilaisten töiden tehostamisessa. Tekoälyn mahdollinen väärinkäyttö luo kuitenkin merkittäviä riskejä, joita mallien kehittäjien sekä päättäjien tulee ottaa huomioon. Tärkeä kysymys onkin, miten tekoälyn tuoma potentiaali saadaan hyödynnettyä siten, että sitä ei pystytä käyttämään pahantahtoisiin käyttötarkoituksiin.

Lisätietoja

Kauppatieteiden kandidaatti, alikersantti Aleksi Simola on Puolustusvoimien tutkimuslaitoksen informaatiotekniikkaosaston tutkimusavustaja. Yhteyshenkilönä Puolustusvoimien tutkimuslaitoksen informaatiotekniikkaosaston vanhempi tutkija TkT Mika Helsingius (p. 0299 800)

[33] Ibid.
[34] O. Varol, E. Ferrara, C. Davis A., F. Menczer ja A. Flammini, ”Online Human-Bot Interactions: Detection, Estimation, and Characterization,” *Proceedings of the international AAAI conference on web and social media*, Montreal, 2017.
[35] E. Ferrara, O. Varol, C. Davis, F. Menczer ja A. Flammini , ”The Rise of Social Bots,” *Communications of the ACM,* osa/vuosik. 59, nro 7, s. 96–104, 2016.
[36] Pew Research Center, “Social Media and News Fact Sheet.” Pewresearch.org. Luettu: 04.11.2024. [Verkossa.] Saatavilla: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/
[37] N. Newman, R Fletcher, C.T. Robertson, A. Ross Arguedas ja R.K. Nielsen. “Reuters Institute Digital News Report 2024.” Reuters Institute for the Study of Journalism, 2024.
[38] A. Bessi ja E. Ferrara, ”Social Bots Distort the 2016 US Presidential Election Online Discussion,” *First Monday,* osa/vuosik. 21, nro 11, 2016.
[39] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn ja M. Dredze, ”Weaponized Health Communication: Twitter Bots,” *American Journal of Public Health,* osa/vuosik. 108, nro 10, s. 1378–1384, 2018.
[40] M. Zhang, X. Qi, Z. Chen ja J. Liu, ”Social Bots’ Involvement in the COVID-19 Vaccine Discussions on Twitter,” *International Journal of Environmental Research and Public Health,* osa/vuosik. 19, nro 3, s. 1651, 2022.
[41] Sedova ym. (2021)
[42] Sedova ym. (2021)