Karelia-ammattikorkeakoulu
Medianomi (AMK)

Deepfake-ilmiön käsitteleminen mediakasvatusvideolla

Joonas Jutila

Opinnäytetyö, marraskuu 2022

www.karelia.fi

Karelia AMMATTIKORKEAKOULU

OPINNÄYTETYÖ
Marraskuu 2022
Media-alan koulutus

Tikkarinne 9
80200 JOENSUU
+358 13 260 600 (vaihde)

Tekijä(t)
Joonas Jutila

Nimeke
Deepfake-ilmiön käsitteleminen mediakasvatusvideolla

Toimeksiantaja
Kansallinen audiovisuaalinen instituutti

Tiivistelmä

Tässä opinnäytetyössä käsitellään deepfake-ilmiötä eli median syväväärennöksiä misinformaatioon ja disinformaatioon liittyvänä ilmiönä. Tavoitteena oli tuottaa tietoa, jolla edistetään kriittistä medialukutaitoa auttamalla ihmisiä ymmärtämään, tunnistamaan ja käsittelemään syväväärennöksiä medioissa. Opinnäytetyön toiminnallisen osan tehtävänä oli tuottaa deepfake-ilmiötä käsittelevä lapsille ja nuorille suunnattu tiivis mediakasvatusvideo. Työn toimeksiantajana on Kansallisen audiovisuaalisen instituutin mediakasvatus- ja kuvaohjelmayksikkö.

Opinnäytetyön tietoperusta koostuu kirjallisuudesta ja verkkolähteistä. Tietoperusta käsittelee tekoälyn avulla manipuloitua media-aineistoa eli syväväärennöksiä mediayhteiskunnassa kohdattavana melko uutena ilmiönä. Työssä pohditaan syväväärennösten uhkakuvia, mahdollisuuksia sekä keinoja uhkien torjuntaan. Toiminnallisena osana tehty mediakasvatusvideo toteutettiin kaksiulotteisena liikegrafiikka-animaationa, joka sisältää videomateriaalia myös omasta esimerkkiväärennöksestäni ja sen tekoprosessista.

Työn tuloksena syntyi syväväärennöksiä käsittelevä kahden minuutin pituinen mediakasvatusvideo sekä monipuolinen tietoperusta. Valmis mediakasvatusvideo on julkaistu Asiaa mediakasvatuksesta -YouTube-kanavalla ja Mediataitokoulun verkkosivustolla.

Kieli
suomi

Sivuja 43
Liitteet 1
Liitesivumäärä 1

Asiasanat
disinformaatio, tekoäly, neuroverkot, manipulaatio, medialukutaito

Karelia
UNIVERSITY OF APPLIED SCIENCES

THESIS
November 2022
Degree Programme in Media

Tikkarinne 9
FI 80200 JOENSUU
FINLAND
Tel. +350 13 260 600

Author(s)
Joonas Jutila

Title
Describing the Deepfake Phenomenon via an Educational Video

Commissioned by
National Audiovisual Institute

Abstract

This thesis discusses deepfakes, or synthetic media, as a phenomenon related to misinformation and disinformation. The goal was to produce information that promotes critical media literacy by helping people understand, identify, and deal with deepfakes in the media. The aim of the practical part of the thesis was to produce a concise educational video about deepfakes, aimed at children and pupils. The work is commissioned by the National Audiovisual Institute's Department for Media Education and Audiovisual Media.

The theoretical background of the thesis consists of literature and web sources. The thesis studies deepfakes, a common term for media that have been manipulated using artificial intelligence, as a recent phenomenon encountered in the media society. The thesis considers the threats and possibilities related to deepfakes, as well as ways to combat these threats. The educational video was realised as a two-dimensional motion graphics animation which also contains video material from my own example deepfake and its creation process.

The work resulted in a two-minute-long educational video and a versatile knowledge base examining deepfakes. The completed educational video has been published on the Asiaa mediakasvatuksesta YouTube channel and on the website of Mediataitokoulu.

Language	Pages 43
Finnish	Appendices 1
	Pages of Appendices 1

Keywords
disinformation, artificial intelligence, neural networks, manipulation, media literacy

Sisältö

Sanasto

Algoritmi
Lista ohjeita, joita tietokone seuraa täsmällisesti (Siltanen 2018).

Autoenkooderimenetelmä
Syväväärennösten luomistapa, jossa ensimmäinen algoritmi pelkistää ja tiivistää sille syötetyn materiaalin, ja toinen algoritmi pyrkii luomaan tiivistetystä datasta alkuperäistä aineistoa vastaavaa materiaalia (Nguyen ym. 2019, 1–2; Sample 2020).

Cheapfake, shallowfake, pintaväärennös
Syväväärennöstä yksinkertaisempi, ilman tekoälyä luotu tai manipuloitu media-aineisto (Ajder, Patrini, Cavalli & Cullen 2019, 11; Hao 2019).

Deepfake, syväväärennös, synteettinen media
Tekoälyn syväoppimisen avulla luotu tai manipuloitu media-aineisto (Paris & Donovan 2019, 2; Somers 2020).

Dekooderi
Autoenkooderimenetelmässä käytettävä algoritmi, joka pyrkii palauttamaan enkooderin pelkistämän materiaalin alkuperäiseen muotoonsa (Nguyen ym. 2019, 1–2; Sample 2020).

Disinformaatio
Tahallisesti levitetty väärä tieto (Schick 2020, 9–12).

Diskriminaattori
Generative Adversarial Network -menetelmässä käytettävä algoritmi, joka pyrkii tunnistamaan generaattorin luomat väärennökset aidon lähdemateriaalin joukosta (Chesney & Citron 2019, 1759–1760; Schick 2020, 44–45).

Enkooderi
Autoenkooderimenetelmässä käytettävä algoritmi, joka pyrkii pelkistämään ja tiivistämään sille syötettyä lähdemateriaalia (Nguyen ym. 2019, 1–2; Sample 2020).

Generaattori
Generative Adversarial Network -menetelmässä käytettävä algoritmi, joka pyrkii luomaan lähdemateriaalia muistuttavaa väärennettyä materiaalia (Chesney & Citron 2019, 1759–1760; Schick 2020, 44–45).

Generative Adversarial Network, GAN
Kahdesta toisiaan vastaan toimivasta algoritmista koostuva neuroverkko, jossa ensimmäinen algoritmi pyrkii luomaan synteettistä aineistoa ja toinen puolestaan tunnistamaan synteettisen aineiston aidon lähdemateriaalin joukosta. Algoritmien oppiessa ne myös kehittävät toinen toistaan. (Chesney & Citron 2019, 1759–1760; Schick 2020, 44–45.)

Misinformaatio
Tahattomasti levitetty väärä tieto (Schick 2020, 9–12).

Neuroverkko
Oppimisalgoritmi, joka pyrkii tuottamaan pyydetyn lopputuotteen sille syötetyn datan perusteella (Hallamaa 2018).

1 Johdanto

Opinnäytetyössä käsitellään deepfake-ilmiötä eli median syväväärennöksiä misinformaatioon ja disinformaatioon liittyvänä ilmiönä mediayhteiskunnassa. Aihe on varsin ajankohtainen ja tärkeä, sillä tekoälyn avulla luodut syväväärennökset ovat vielä verrattain tuore ilmiö, josta kaikkien nykyaikaisessa mediayhteiskunnassa elävien olisi hyvä olla tietoisia. Vaikka median manipulaatiota on tehty ennenkin, tuo deepfake-teknologia mukanaan uudenlaisia uhkia, joihin moni ei välttämättä vielä osaa varautua. Uudenlaisten uhkien edessä tarvitaan mediakasvatusta sekä kriittistä medialukutaitoa.

Opinnäytetyön tietoperusta pohjautuu enimmäkseen englanninkieliseen kirjallisuuteen ja verkkoartikkeleihin sekä joihinkin suomenkielisiin verkkolähteisiin. Aiheen käsittely suomalaisessa tietokirjallisuudessa ei vielä ole ollut kovin laajamittaista. Tietoperustassa perehdytään deepfake-ilmiön historiaan, teknologisiin perustietoihin, uhkakuviin, hyötyihin sekä keinoihin, joilla väärennöksiä voidaan tunnistaa ja väärennösten uhkia voidaan torjua. Lisäksi opinnäytetyön tietoperustassa pohditaan syväväärennösten tekemiseen liittyviä kysymyksiä etiikasta sekä henkilötietojen käsittelystä. Tietoperusta on pyritty kirjoittamaan siten, että sen pystyy ymmärtämään ilman syvällistä teknologista perehtymistä.

Työn toiminnallisessa osassa tuotettiin syväväärennöksiä käsittelevä tiivis mediakasvatusvideo, jolla pyritään tukemaan lasten ja nuorten kriittistä medialukutaitoa. Mediakasvatusvideon tavoitteena on auttaa nuoria ymmärtämään, mitä syväväärennökset ovat sekä antaa heille keinoja mediassa kohtaamiensa väärennösten tunnistamiseen. Video tehtiin toimeksiantona Kansallisen audiovisuaalisen instituutin mediakasvatus- ja kuvaohjelmayksikölle osana Uudet lukutaidot -kehittämisohjelmaa. Mediakasvatusvideo koostuu Faceswap-ohjelmistolla tuottamastani syväväärennöksestä sekä kaksiulotteisesta liikegrafiikasta. Omalla syväväärennökselläni pyrin myös osoittamaan, ettei väärennöksen tekemiseen vaadita erityisen laajaa teknistä perehtymistä.

2 Mitä deepfaket ovat?

2.1 Deepfake-ilmiöön liittyviä käsitteitä

Termi deepfake tulee englannin kielen sanoista ”deep learning” ja ”fake”, jotka tarkoittavat vastaavasti syväoppimista ja väärennöstä. Käsitteenä deepfake tarkoittaakin tekoälyn syväoppimisen avulla luotua tai manipuloitua videota, kuvaa, ääntä tai tekstiä. Tekoälyn avulla luotua tai manipuloitua media-aineistoa kutsutaan myös synteettiseksi eli keinotekoiseksi mediaksi. (Paris & Donovan 2019, 2; Somers 2020.) Yleisradion teknologiatoimittaja Teemu Hallamaa (2019) on suomentanut deepfake-termin syväväärennökseksi. Tämä suomennos on lisätty myös Kotimaisten kielten keskuksen sanatietokantaan vuonna 2019 (Eronen ym. 2019). Suomennos vaikuttaa luontevalta, joten sitä käytetään myös tässä opinnäytetyössä. Suurin osa tähän mennessä tehdyistä syväväärennöksistä on videoita, joissa alun perin esiintyneiden henkilöiden kasvojen tilalle on vaihdettu toisen henkilön kasvot (Ajder ym. 2019). Tekoälyn avulla voidaan kuitenkin luoda myös vakuuttavan kuuloista synteettistä puhetta (Chesney & Citron 2019, 1757), joskin se on teknisesti videon manipulointia haastavampaa (Hallamaa 2021).

Syväväärennösten asiayhteydessä mainitaan usein myös joko termi cheapfake tai shallowfake. Näillä kummallakin tarkoitetaan syväväärennöstä yksinkertaisempaa, ilman tekoälyä luotua median manipulaatiota (Ajder ym. 2019, 11; Hao 2019). Cheapfake-termi saattaa kuitenkin antaa liian vaarattoman mielikuvan halvasta väärennöksestä, joten tässä opinnäytetyössä näistä väärennöksistä käytetään termiä shallowfake, jonka voisi suomentaa pintaväärennökseksi. Käytännössä termeillä tarkoitetaan esimerkiksi kaksoisolentojen kuvaamista tai imitaattorin äänittämistä, median nopeuttamista tai hidastamista, yksityiskohtien manipuloimista editointiohjelmistolla tai yksinkertaisesti aidon median käyttöä harhaanjohtavassa kontekstissa (Ajder ym. 2019, 11; Paris & Donovan 2019, 6, 14–16). Vaikka pintaväärennökset saattavat olla teknisesti yksinkertaisia, ovat ne silti tehokkaita työkaluja ihmisten harhaanjohtamiseen (Paris & Donovan 2019, 14–16; Sample 2020).

2.2 Median manipulaation historiaa

Mediasisältöjen manipulaatio ei ole uusi ilmiö, sillä valokuvia on manipuloitu jo 1800-luvun loppupuolelta saakka. Esimerkiksi Josef Stalinin tiedetään poistattaneen 1930-luvun puhdistuksissa menehtyneitä poliittisia vastustajiaan tätä aiemmin otetuista valokuvista. Tuolloin kuvien manipuloimiseen vaadittiin vielä hyvin tarkkaa käsityötä. (Schick 2020, 26–29.) Teknologian kehitys on kuitenkin johtanut siihen, että median manipuloinnista on tullut vaivattomampaa kuin koskaan aiemmin. Adobe Photoshop ja sitä vastaavat kuvankäsittelyohjelmistot ovat modernisoineet kuvien muokkaamisen 1990-luvun alusta lähtien. (Somers 2020.) Nykyisin valokuvien, videoiden ja äänen muokkaamiseen ja manipuloimiseen on saatavilla runsaasti erilaisia sovelluksia niin tietokoneille kuin mobiililaitteillekin.

Ensimmäiset syväväärennökset on jäljitetty syksyyn 2017 ja Reddit-yhteisöpalvelun käyttäjään ”u/deepfakes”. Hänen lataamansa syväväärennökset olivat pornografisia videoita, joissa esiintyneiden naisten kasvojen tilalle oli vaihdettu tunnettujen naispuolisten julkisuuden henkilöiden kasvot. Nämä ensimmäiset syväväärennökset on sittemmin poistettu Redditistä. (Ajder ym. 2019, 3; Schick 2020, 35.) Väärennösten tekijä, ”u/deepfakes”, ehti kuitenkin jakaa deepfake-algoritminsa lähdekoodin vapaasti saataville, jotta muutkin voisivat kehittää omia deepfake-ohjelmistojaan sen pohjalta (Schick 2020, 38).

Internetissä tunnistettujen syväväärennösten määrä on kaksinkertaistunut noin kuuden kuukauden välein vuoden 2018 lopulta alkaen. Tunnistettujen deepfake-videoiden määrä yli kymmenkertaistui alle 8 000 kappaleesta 85 000 kappaleeseen vuosien 2019 ja 2020 aikana. (Cavalli 2021.) Tämän jälkeen tunnistettujen syväväärennösten määrästä ei ole julkaistu vertailukelpoisia tilastoja, mutta tekemäni Google Trends -haun mukaan ainakin deepfake-hakusanan keskimääräinen käyttöaste on jatkanut kasvuaan myös vuosien 2021 ja 2022 aikana (kuvio 1). Tähän mennessä suurin hetkellinen kiinnostus hakusanaa kohtaan on kuitenkin havaittu jo maaliskuussa 2021. (Google Trends 2022a.) Tuolloin aiheeseen liittyvissä hakutuloksissa korostuivat näyttelijä Tom Cruisesta tehdyt, mediassakin laajalle levinneet syväväärennökset (Google Trends 2022b).

Alueellisesti suurin suhteellinen kiinnostus deepfake-hakusanaa kohtaan on tähän mennessä ollut Etelä-Koreassa (Google Trends 2022a).

Kuvio 1. Deepfake-hakusanan käyttöaste Googlessa (Google Trends 2022a).

Vielä vuonna 2019 noin 96 prosenttia internetissä olevista syväväärennöksistä oli pornografisia videoita (Ajder ym. 2019, 1). Pornografian lisäksi syväväärennöksiä on tehty myös esimerkiksi viihde- ja opetuskäyttöön. Näitä humoristisia ja opettavaisia deepfake-videoita on tehty niin näyttelijöistä, poliittisista toimijoista kuin yritysmaailman henkilöistäkin. (Parkin 2019; Sample 2020.) Julkisuuden henkilöt valikoituvat usein väärennösten kohteiksi, koska heistä on ladattu runsaasti kuva- ja videomateriaalia internetiin (Hallamaa 2019; Nguyen ym. 2019, 1). Tulevaisuudessa syväväärennöksiä saatetaan nähdä enenevissä määrin myös esimerkiksi mainonnassa, poliittisessa vaikuttamisessa sekä rahallisissa huijauksissa.

2.3 Syväväärennösten tekoprosessi

Videon syväväärentäminen on nykyisin suhteellisen helppoa, kun väärennösohjelmistot lähdekoodeineen ovat levinneet yleisesti saataville. Synteettistä puhetta tuottavat algoritmit sen sijaan eivät vielä ole levinneet samalla tavalla yleiseen käyttöön. (Hallamaa 2021.) Videomanipulaation tekeminen koostuu kolmesta työvaiheesta, jotka ovat:

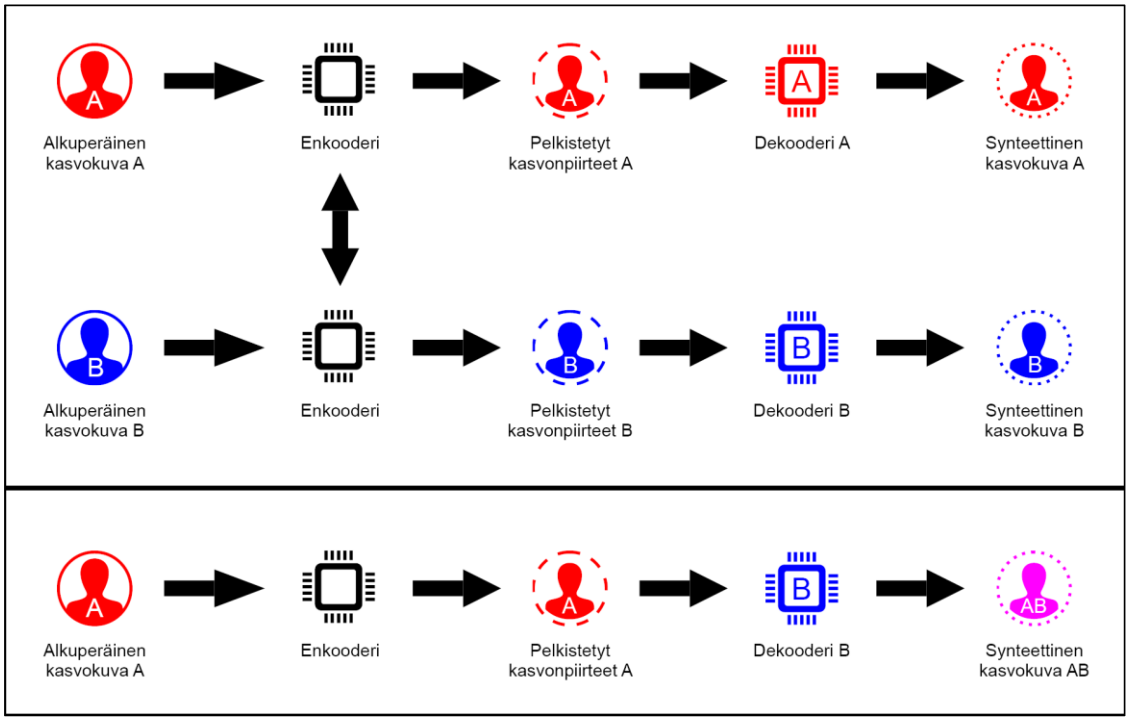
- kasvojen tunnistaminen ja paikantaminen lähdemateriaalista
- tekoälyalgoritmien kouluttaminen synteettisten kasvokuvien luomiseen
- synteettisten kasvokuvien siirtäminen lähdemateriaaliin. (Das ym. 2021, 80–81.)

Deepfake-ohjelmistot hyödyntävät tekoälyn neuroverkkoja, jotka ovat eräänlaisia oppimisalgoritmeja. Neuroverkoille syötetään runsaita määriä dataa, jonka perusteella ne pyrkivät itsenäisesti luomaan pyydettyjä lopputuotteita, tässä tapauksessa siis keinotekoista media-aineistoa. Neuroverkkojen kehitys on nopeutunut viime vuosina erityisesti tietokoneiden grafiikkasuorittimien käytön vuoksi. Perinteisillä tietokoneprosessoreilla aineisto pitäisi käydä läpi tavu kerrallaan, kun grafiikkasuorittimet taas kykenevät kymmeniin tuhansiin samanaikaisiin laskutoimituksiin, mikä nopeuttaa neuroverkkojen kykyä oppia aineistoa. (Hallamaa 2018.)

Neuroverkkojen kouluttaminen on videomanipulaation laajin työvaihe. Se on toistuva prosessi, johon tarvitaan yleensä vähintään satoja tai tuhansia kasvokuvia kahdesta eri henkilöstä sekä riittävästi aikaa prosessin kymmeniin tuhansiin toistokertoihin. (Hallamaa 2019; Schick 2020**,** 44–45.) Syväväärennöksen tekemiseen käytettävien kasvokuvien ei välttämättä tarvitse edes muistuttaa toisiaan, vaan tärkeämpää saattaa olla kasvojen eleiden selkeys (Hao 2020). Neuroverkkojen kouluttamiseen käytettyjen kasvokuvien määrä, monipuolisuus, kuvanlaatu sekä kouluttamiseen käytetty aika vaikuttavat syväväärennöksen lopulliseen laatuun. Tietokoneen suorituskyky ja syväväärennöksen haluttu kuvatarkkuus puolestaan vaikuttavat prosessin ajalliseen kestoon, mutta tehokkaallakin tietokoneella prosessiin kuluu useita tunteja. (Kim ym. 2018, 6; Korshunov & Marcel. 2020, 3.) Koulutusprosessin nopeuttamiseksi on mahdollista hyödyntää internetin pilvilaskentapalveluita, jotka mahdollistavat laskentatehon vuokraamisen edulliseen hintaan (Das ym. 2021, 14).

Ensimmäinen tapa tekoälyn kouluttamiseen on niin sanottu autoenkooderimenetelmä (kuvio 2). Se pitää sisällään kaksi tekoälyn algoritmiparia: kahden henkilön kasvokuvien jakamat enkooderit ja molemmille omat dekooderinsa. Jaetuille enkoodereille syötetään satoja kasvokuvia kahdesta henkilöstä, minkä

jälkeen ne pyrkivät tunnistamaan ja oppimaan kasvokuvien välillä olevia yhtäläisyyksiä sekä pelkistämään ja tiivistämään kuvat yhteisiin piirteisiinsä. Dekooderit puolestaan oppivat luomaan enkooderien pelkistämistä kuvista oman henkilönsä alkuperäisiä kasvokuvia muistuttavia synteettisiä kuvia. Lopuksi enkooderien pelkistämät kuvat syötetään toiselle dekooderille, joka luo uudet kasvot, joissa ensimmäisen henkilön kasvonpiirteet ja eleet on siirretty jälkimmäisen henkilön kasvoille. Videomanipulaatiossa tämä prosessi tehdään jokaiselle yksittäiselle kuvaruudulle. (Nguyen ym. 2019, 1–2; Sample 2020.)



Kuvio 2. Autoenkooderimenetelmä havainnollistettuna (Nguyen ym. 2019, 3).

Toinen tekoälyn koulutustapa on GAN-menetelmä, joka tulee sanoista ”Generative Adversarial Network”. Se koostuu kahdesta toisiaan vastaan toimivasta algoritmista: generaattorista ja diskriminaattorista. Molemmille algoritmeille syötetään sama lähdemateriaali, minkä jälkeen ensimmäinen algoritmi pyrkii luomaan väärennettyä mediaa, ja toinen puolestaan tunnistamaan väärennökset aidon lähdemateriaalin joukosta. Koulutusprosessin edetessä generaattori oppii luomaan jatkuvasti tarkempia väärennöksiä, kun taas diskriminaattori oppii tunnistamaan yhä vakuuttavampia väärennöksiä. Näin algoritmit kehittävät toisiaan, ja lopputuloksena syntyvästä syväväärennöksestä syntyy mahdollisimman

aidonnäköinen tai -kuuloinen. (Chesney & Citron 2019, 1759–1760; Schick 2020, 44–45.)

Teoriassa GAN-menetelmän toisiaan kehittävät algoritmit johtavat lopulta täydellisiin syväväärennöksiin. Joidenkin arvioiden mukaan olemme enää muutaman vuoden päässä täydellisistä syväväärennöksistä, joiden tunnistaminen voi osoittautua mahdottomaksi. (Schick 2020, 45–50.) Tällä hetkellä esimerkiksi silmälasit saattavat tuottaa tekoälylle haasteita kasvokuvien yhdistämisessä (Hao 2020). Teknologian kehittyessä syväväärennöksistä tulee kuitenkin entistä laadukkaampia samalla kun niiden tekeminen helpottuu, nopeutuu ja tulee vaatimaan entistä vähemmän resursseja (Schick 2020, 45–50).

Tähän mennessä suosituimmat kaksi deepfake-ohjelmistoa ovat DeepFaceLab ja Faceswap (Schick 2020, 38–39). Kumpikin ohjelmisto on ladattavissa ilmaiseksi GitHub-ohjelmistonkehityspalvelusta. Ne molemmat perustuvat ensimmäisenä esiteltyyn autoenkooderimenetelmään (Nguyen ym. 2019, 2), mutta DeepFaceLab mahdollistaa käyttöohjeidensa mukaan myös GAN-menetelmän hyödyntämisen (TMBDF 2022). DeepFaceLab-ohjelmisto vaikuttaa kuitenkin eettisesti kyseenalaiselta, sillä sen käyttöohjeet ja keskustelupalsta sijaitsevat deepfake-pornosivustolla (TMBDF 2022). Deepfake-pornografian ongelmallisuudesta kerrotaan tarkemmin opinnäytetyön luvussa 3.1. Faceswap-ohjelmistolla puolestaan on oma verkko-osoitteensa (Faceswap 2022) sekä julistus ohjelmiston eettisestä käytöstä (Deepfakes 2021). Lisäksi ohjelmistossa on helppokäyttöinen graafinen käyttöliittymä. Tässä esitettyjen syiden vuoksi opinnäytetyön toiminnallisessa osassa käytetään Faceswapia.

DeepFaceLabin pohjalta on luotu myös DeepFaceLive-ohjelmisto, jolla kyetään väärentämään reaaliaikaista videokuvaa. Ohjelmistossa hyödynnetään julkisuuden henkilöiden kuvilla valmiiksi koulutettuja algoritmeja. Koska algoritmia ei enää erikseen kouluteta loppukäyttäjän kasvokuvilla, vakuuttavan väärennöksen luominen vaatii jonkin verran yhdennäköisyyttä henkilöiden kasvojen välillä. Reaaliaikaisilla syväväärennöksillä on mahdollista väärentää esimerkiksi videopuheluita, joissa usein liikutaan kohtalaisen vähän. (Anderson 2021.)

Edellä esiteltyjen väärennöstapojen lisäksi markkinoille on saapunut useita mobiilisovelluksia, joiden avulla tavalliset kuluttajat voivat tehdä omia syväväärennöksiään jopa yksittäisestä kuvasta. Sovellusten käyttämä teknologia ei vielä ole yhtä kehittynyttä kuin vastaavien tietokoneohjelmistojen, mutta ne ovat huomattavasti helppokäyttöisempiä. Tämän lisäksi osa mobiilisovelluksista mahdollistaa vain ennalta määriteltyjen, harmittomampien kohdevideoiden väärentämisen. (Fowler 2021.) Teknologia kuitenkin kehittyy nopeasti, ja tulevaisuudessa näitä mobiilisovelluksiakin saattaa olla mahdollista käyttää entistä monipuolisempiin tarkoituksiin. Teknologinen kehitys tuo mukanaan myös uudenlaisia uhkia.

3 Syväväärennösten uhkakuvat

3.1 Uhka yksityisyydelle

Eräs syväväärennösten konkreettisimmista tähänastisista haitoista on niiden luoma uhka yksittäisten henkilöiden yksityisyydelle. Syväväärennösten avulla lähes kenen vain voidaan nimittäin esittää sanoneen tai tehneen mitä tahansa. Tämä uhka koskee kaikkia, jotka esiintyvät internetiin ladatuissa valokuvissa, videoissa tai äänitallenteissa. (Schick 2020, 148–149.) Jo minuutin pituisesta videosta on mahdollista poimia satoja erilaisia kasvokuvia tekoälyn kouluttamiseen.

Syväväärennösten yksityisyydelle luoma uhka on nykyisellään erittäin sukupuolittunut ongelma (Chesney & Citron 2019, 1773). Suurin osa internetiin ladatuista syväväärennöksistä on ollut pornografista sisältöä (Ajder ym. 2019, 1), ja deepfake-pornon uhrit puolestaan ovat suurelta osin naispuolisia. Ongelmaa kuvaa se, että naisiin kohdistuvaa deepfake-pornoa on tehty niin paljon, että sitä on alettu luokittelemaan internetissä eri kategorioihin, kun taas miehet ovat enimmäkseen välttyneet deepfake-pornon uhriksi joutumiselta. (Schick 2020, 158–160.) Ongelmaa on kuvattu muun muassa sukupuolittuneeksi riistoksi (Chesney & Citron 2019, 1773) sekä kuviin pohjautuvaksi väkivallaksi (Paris &

Donovan 2019, 26). Pornografiset väärennökset aiheuttavat uhreilleen usein kärsimystä ja häpeää, vaikka ne myöhemmin todettaisiinkin väärennöksiksi. Tässä vaiheessa väärennökset ovat usein jo levinneet laajalle aiheuttaen henkilökohtaista vahinkoa. (Chesney & Citron 2019, 1774.)

Syväväärennösten tekeminen tai levittäminen ei itsessään ole laitonta, mutta deepfake-pornon tekijät syyllistyvät usein sekä tekijänoikeusrikkomukseen että uhrin kunnianloukkaukseen (Chesney & Citron 2019, 1788, 1793; Sample 2020). Lisäksi henkilön kasvokuvia pidetään Euroopan unionin yleisen tietosuoja-asetuksen eli GDPR:n alaisina henkilötietoina, jolloin kasvokuvien käsittely edellyttää laillista käsittelyperustetta, kuten kohdehenkilön suostumusta tai esimerkiksi sananvapauteen perustuvaa oikeutettua etua (Das ym. 2021, 38–39). Väärennetyt pornovideot kuitenkin ladataan internetiin usein joko kokonaan anonyymisti tai nimimerkin takaa, jolloin pahantahtoisten videoiden tekijät voi olla haastavaa saada vastaamaan teoistaan (Chesney & Citron 2019, 1792). Hyväntahtoistakin syväväärennöstä tehdessä tulisi pohtia eettisiä kysymyksiä, kuten kenestä väärennöksen tekeminen on sopivaa, ja miten väärennetty media-aineisto tulee merkitä syväväärennökseksi. Monet yksityishenkilöt tai erinäiset pahantahtoiset toimijat eivät kuitenkaan välttämättä välitä eettisistä tai tekijänoikeudellisista rajoitteista.

3.2 Taloudellinen uhka

Syväväärennökset uhkaavat yksityishenkilöiden lisäksi myös yrityksiä. Otolliseen aikaan julkaistu yritystä koskeva negatiivinen syväväärennös voisi aiheuttaa mittavia mainehaittoja. Tämä taas voisi johtaa esimerkiksi yrityksen osakekurssin laskuun tai yhteistyösopimusten purkautumisiin, jolloin väärennöksestä koituisi myös taloudellisia vaikutuksia. (Das ym. 2021, 31.) Tällaisia mainesabotaasiin pyrkiviä väärennöksiä ei vielä ole juurikaan nähty, mutta myös niihin tulee varautua jatkossa (Chesney & Citron 2019, 1774–1777).

Mainesabotaasin sijaan väärentäjät ovat jo pyrkineet yritysten rahalliseen huijaamiseen, osa heistä myös onnistuen tavoitteessaan. Syväväärennösten avulla

on huijattu yrityksiä ja varakkaita yksityishenkilöitä yhteensä jo kymmenien miljoonien eurojen edestä. (Schick 2020, 139–140, 145–146.) Huijarit ovat esimerkiksi väärentäneet yrityksen toimitusjohtajan ääntä ja pyytäneet yrityksessä työskenteleviä toimihenkilöitä siirtämään yrityksen varoja ilmoittamalleen pankkitilille (Ajder ym. 2019, 14). Edellä mainitut kymmenien miljoonien menetykset eivät vaikuta laajemmassa mittakaavassa vielä erityisen suurilta, mutta ne ovat kuitenkin merkki siitä, että sekä yritysten että yksityishenkilöiden tulee huolehtia medialukutaidostaan ja tietoturvallisuudestaan.

Tekoälyä hyödyntävä varmennus- ja tunnistuspalvelu Sensity raportoi, että heidän suorittamassaan testissä syväväärennökset läpäisivät useimmat finanssialalla käytettävät asiakkaantunnistusprosessit ongelmitta. Testissä kyettiin väärentämään sekä henkilökorttien kasvokuvia että reaaliaikaista videokuvaa, jossa henkilö noudattaa tunnistusprosessin ohjeita esimerkiksi sulkemalla silmänsä pyydettäessä tai kääntämällä päätään ohjeissa toivottuun suuntaan. Raportissa mainitaan kymmenten jopa miljoonia asiakkaita palvelevien yritysten turvautuvan vastaaviin, huijattavissa oleviin asiakkaantunnistusprosesseihin. Tämä altistaa yritykset suuremmillekin rahallisille petoksille. (Sensity 2022.)

Myös FBI:n verkkorikollisuuden raportointikeskus IC3 on varoittanut yrityksiä syväväärennöksistä. IC3:n varoitus koskee syväväärennösten kasvanutta käyttöä tietotekniikka-alan etätyönhaussa. Pahantahtoiset toimijat voivat käyttää reaaliaikaisia video- tai äänimanipulaatioita työhaastatteluissa päästäkseen etätyötehtäviin, joissa heillä on pääsy yritysten tietoverkkoihin tai asiakkaiden henkilö- tai taloustietoihin. Onnistuessaan tällainen huijaus voisi tuottaa vahinkoa sekä kohteena olevalle yritykselle että sen asiakkaille. (Internet Crime Complaint Center (IC3) 2022.)

3.3 Uhka demokratialle

Syväväärennökset voivat aiheuttaa vahinkoa myös laajemmalle yhteiskunnalle. Median väärennöksillä voidaan nimittäin pyrkiä heikentämään yleistä

luottamusta sekoittamalla valhe ja totuus, sekä syventämään yhteiskunnan polarisaatiota eli kahtiajakoa. (Chesney & Citron 2019, 1777–1778; Jantunen 2016, 2–3.)

Geopolitiikan, teknologian ja kyberturvallisuuden asiantuntija Nina Schick (2020, 9–12) antaa heti syväväärennöksiin keskittyvän kirjansa alussa pessimistisen kuvan 2020-luvun alun tietoympäristöstä. Hänen mukaansa länsimainen tietoyhteiskunta on jo entuudestaan rikki ja altis hyväksikäytölle, kun misinformaatiota ja disinformaatiota eli tahattomasti ja tahallisesti levitettyä väärää tietoa on liikkeellä enemmän kuin koskaan aiemmin. Syväväärennökset edistävät tätä negatiivista kehitystä vielä entisestään. (Schick 2020, 9–12.)

Myös ihmisten kognitiiviset harhat sekä sosiaalisen median alustojen toimintaperiaatteet edistävät väärän tiedon leviämistä yhteiskunnassamme. Ihmiset ovat taipuvaisia uskomaan omia ennakkokäsityksiään vahvistavaa tietoa, ja sosiaalisen median algoritmitkin pyrkivät suodattamaan ihmisille näyttämäänsä sisältöä vastaamaan heidän omaa ajatusmaailmaansa. Tämän lisäksi negatiivinen ja väärä tieto pyrkii usein vaikuttamaan tunteisiin, jolloin se jää paremmin ihmisten mieliin ja tulee jaetuksi entistä laajemmalle yleisölle. Tämä kaikki syventää yhteiskunnan polarisaatiota ja heikentää yleistä luottamusta. (Chesney & Citron 2019, 1763–1768.)

Yhteiskunnan polarisaatio on huolestuttava ilmiö, joka vaikeuttaa yhteiskunnallista keskustelua. On haastavaa käydä järkevää keskustelua, jos keskustelun osapuolet elävät omissa kuplissaan, joissa molemmilla on eri näkemys totuudesta. (Schick 2020, 12–13.) Disinformaatiovaikuttaminen perustuukin usein juuri kansan harhauttamiseen ja kahtiajakoon. Järkevien ja perusteltujen päätösten tekeminen hankaloituu, kun totuus ja valhe sekoittuvat. (Schick 2020, 54.)

Syväväärennöksillä voidaan myös pyrkiä horjuttamaan instituutioiden nauttimaa luottamusta, hidastamaan uutistoimitusten työtä sekä vaikuttamaan esimerkiksi vaalituloksiin (Chesney & Citron 2019, 1778–1785). Yhdysvaltain vuoden 2020 presidentinvaalit koettiin etukäteen yhdeksi syväväärennöksillä vaikuttamisen

suurimmista vaaranpaikoista (Hao 2019; Parkin 2019), mutta tällaista ei lopulta havaittu. Sen sijaan vaalituloksesta on levitetty runsaasti perinteisempää mis- ja disinformaatiota (Gerhart 2021). Ylipäätään emme vielä ole nähneet syväväärennöksiä käytettävän laajemman mittakaavan yhteiskunnalliseen vaikuttamiseen, eivätkä kaikki suhtaudu uhkaan yhtä vakavasti. Syväväärennökset ovat ainakin vielä tekoälyn tunnistettavissa, jolloin perinteisempi disinformaatio ja niin kutsutut pintaväärennökset saattavat olla yhtä tehokkaita, mutta edullisempia ja nopeampia keinoja disinformaation levittämiseen. (Brandom 2019; Uchill 2019.) Uhkan mahdollisuutta ei kuitenkaan kannata poissulkea, sillä deepfake-teknologia kehittyy ja yleistyy nopeasti.

Edellä mainittujen yhteiskunnallisten uhkien lisäksi syväväärennökset antavat epärehellisille toimijoille uudenlaisen mahdollisuuden totuuden kiistämiseen. He voivat nyt väittää aitoakin media- tai todistusaineistoa manipuloiduksi, minkä ansiosta he saattavat selvitä itseensä kohdistuneista syytöksistä tai ainakin heikentää niiden uskottavuutta. Tätä kutsutaan englanninkielisessä tutkimuskirjallisuudessa termillä liars dividend, joka tarkoittaa suomennettuna valehtelijan osinkoa. Käytännössä epärehellisillä toimijoilla on siis valta hyökätä kaikkia vastaan, mutta myös kiistää kaikki itseensä kohdistuvat syytökset. Tämä heikentää osaltaan video- ja äänitallenteiden yleisesti koettua todistusvoimaa. (Chesney & Citron 2019, 1758; Schick 2020, 130.)

3.4 Uhka kansalliselle turvallisuudelle

Syväväärennöksillä voidaan pyrkiä vaikuttamaan myös kansalliseen turvallisuuteen. Sekä Puolustusvoimien tutkimuslaitoksen doktriiniosasto (Jantunen 2016) että Yhdysvaltain asevoimien tutkimusorganisaatio DARPA (Turek 2021) ovat todenneet visuaalisen median manipulaation turvallisuusuhaksi. Syväväärennetty media-aineisto esimerkiksi virkavallan väärinkäytöksistä saattaisi nykyisessä poliittisessa ilmapiirissä johtaa väkivaltaisiin mellakoihin (Chesney & Citron 2019, 1780–1781).

Kansallisesta turvallisuusuhasta nähtiin todellinen esimerkki maaliskuussa 2022, kun Ukrainan presidentti Volodymyr Zelenskyi näytti syväväärennetyssä videossa kehottavan Ukrainan kansalaisia antautumaan Venäjän hyökättyä Ukrainaan. Tämä lienee yksi vaarallisimmista tapauksista, mihin syväväärennöksiä on tähän mennessä käytetty. Kyseinen väärennös ei kuitenkaan osoittautunut erityisen tehokkaaksi, sillä se oli melko heikkolaatuinen, ja koska Ukraina oli varautunut informaatiovaikuttamiseen perusteellisesti. Manipuloitu video pystyttiin kiistämään tehokkaasti, ja länsimaiset sosiaalisen median palvelut poistivat manipulaation nopeasti. (Holroyd & Olorunselu 2022.)

Tekoälyn avulla voidaan väärentää kasvokuvien lisäksi myös satelliittikuvia (Eckart 2021). Väärennettyjä satelliittikuvia on mahdollista käyttää esimerkiksi yleisen paniikin lietsomiseen tai sodankäynnissä vastapuolen taktiseen harhauttamiseen. Kehittyneempien valtioiden tiedusteluviranomaiset ovat päteviä tulkitsemaan näkemäänsä mediaa, mutta kehittyvät valtiot saattaisivat tällaisessa tilanteessa tulla ainakin hetkellisesti harhautetuksi. (Chesney & Citron 2019, 1782–1783.) Julkisen yleisön harhaanjohtaminen olisi luultavasti vielä todennäköisempää, sillä satelliittikuvasto lienee heille lähtökohtaisestikin vieraampaa (Eckart 2021).

Venäjä on ollut informaatiosodankäynnin edelläkävijä aina Neuvostoliiton ajoista lähtien, mutta useat muutkin valtiot seuraavat perässä. Esimerkiksi Kiina, Pohjois-Korea ja Iran ovat kuitenkin vielä tähän saakka olleet disinformaation tuottamisessa Venäjää kömpelömpiä. Varsinkin Kiina on keskittynyt enemmän omaan kansaansa vaikuttamiseen, mutta teknologian kehittyessä myös valtioiden välinen informaatiosodankäynti tullee lisääntymään. Jo 70 eri valtiota tuotti disinformaatiota vuonna 2020. (Schick 2020, 83–85.) Disinformaatiokampanjoilla saatetaan pyrkiä vaikuttamaan myös epäsuorasti ei-valtiollisten toimijoiden kautta, sillä niiden toiminnalle ei ole asetettu samanlaisia vastuita ja periaatteita kuin demokraattisille valtioille (Jantunen 2016).

4 Syväväärennösten hyötykäyttö

Syväväärennösten lukuisista uhkakuvista huolimatta niillä voi olla hyödyllisiäkin käyttötarkoituksia. Niitä voidaan hyödyntää esimerkiksi viihteessä, mediakasvatuksessa, mainonnassa sekä positiivisten representaatioiden mahdollistamisessa. Syväväärennösten avulla historialliset ja edesmenneetkin henkilöt on mahdollista saada esiintymään videolla tai äänitallenteella. Tämänkaltaista teknologiaa on käytetty esimerkiksi Rogue One: A Star Wars Story ja Star Wars: Episode VIII - The Last Jedi -elokuvissa. (Chesney & Citron 2019, 1769–1771.) Lisäksi syväväärennösten avulla kyetään nuorentamaan näyttelijöiden kasvokuvia, mikäli heistä löytyy nuorempana kuvattua lähdemateriaalia. Näin 70-vuotias Mark Hamill kykeni esiintymään itseään nuorempana Luke Skywalkerina The Book of Boba Fett -televisiosarjassa. Yhdysvaltalainen viihde- ja media-alan ammattiliitto SAG-AFTRA kuitenkin korostaa, että myös näyttelijöiden digitaalisesti luotuihin esiintymisiin vaaditaan näyttelijän suostumus sekä kohtuullinen korvaus. (Giardina 2022.)

MyHeritage-sukututkimuspalvelu puolestaan mahdollistaa ihmisille heidän edesmenneiden sukulaistensa kasvojen animoinnin jo yksittäisenkin valokuvan perusteella (Fowler 2021). Kuolleiden henkilöiden esittäminen syväväärennösten avulla tosin herättää eettisiä kysymyksiä. Vaikka nykyinen teknologia mahdollistaakin heidän esittämisensä media-aineistossa, ei se välttämättä ole kunnioittavaa tai eettistä, varsinkaan jos vastaavanlaisesta toiminnasta ei ole keskusteltu ennen heidän kuolemaansa. Aihe nousi uutisotsikoihin kesällä 2021, kun ilmeni, että Roadrunner: A Film About Anthony Bourdain -dokumenttielokuvassa oli hyödynnetty edesmenneen Bourdainin syväväärennettyä ääntä. Manipulaation käytöstä ei kuitenkaan mainittu itse elokuvassa, mikä aiheutti närkästystä sosiaalisen median kommentoijien keskuudessa. (Rosner 2021.)

Syväväärennöksistä voi olla merkittävää hyötyä myös ihmisten autonomialle. Äänen syväväärennöksillä voidaan esimerkiksi luoda äänensä menettäneille henkilöille uusi synteettinen ääni. Vuonna 2021 britannialainen tekoälyä hyödyntävä puheäänipalvelu Sonantic loi synteettisen puheäänen äänensä

kurkkusyövän hoidossa menettäneelle näyttelijälle Val Kilmerille. Keinotekoinen ääni luotiin yhteistyössä Kilmerin kanssa käyttäen hänen aikaisempia äänitallenteitaan lähdemateriaalina. (Flynn 2021.)

Tämän lisäksi teknologia mahdollistaa ihmisille heidän omien kasvojensa liittämisen esimerkiksi videopelihahmoihin tai videolla esiintyviin henkilöihin, jolloin he voivat kokea osallistuvansa hahmojen toimintaan laajemmin (Chesney & Citron 2019, 1770–1771). HBO:n Welcome to Chechnya -dokumenttielokuvassa deepfake-teknologiaa on puolestaan käytetty vainottujen seksuaalivähemmistöjen edustajien identiteetin suojeluun. Elokuvassa vähemmistöjen edustajat saattoivat kertoa omat tarinansa paljastamatta identiteettiään tai menettämättä ihmisyyttään, kun heidän kasvonsa voitiin korvata vapaaehtoisten deepfake-näyttelijöiden kasvoilla. (Hao 2020.)

Edellä mainittujen hyötyjen lisäksi syväväärennökset ovat hyödyllinen lisäkeino mediakasvatukseen, sillä ne antavat laajemmalle yleisölle konkreettisen kuvan väärennetyn median vaaroista. Esimerkiksi Yleisradio (Hallamaa 2019; Hallamaa 2021), britannialainen televisiokanava Channel 4 (2020) sekä uutis- ja viihdeyritys BuzzFeed (2018) ovat jo tehneet arvokasta työtä tällä saralla. Myös tämän opinnäytetyön toiminnallisessa osiossa pyritään mediakasvatuksellisuuteen tuomalla ilmiötä nuorten tietoisuuteen.

5 Syväväärennösten tunnistaminen, sääntely ja uhkien torjunta

Syväväärennösten aiheuttamien uhkien torjumiseksi tarvitaan sekä teknisiä että sosiaalisia ratkaisuja. Teknologian kehitystä ei voi peruuttaa, minkä vuoksi on keksittävä uudenlaisia ratkaisuja. Myös kriittisen medialukutaidon merkitys kasvaa, kun syväväärennösten tunnistaminen vaikeutuu teknologisen kehityksen myötä. (Das ym. 2021, 57–72.)

Osan tämänhetkisistä syväväärennöksistä voi vielä tunnistaa paljaalla silmällä tai korvallakin. Videolla esiintyvän henkilön iho saattaa näyttää ikään kuin muoviselta ja liian huolitellulta tai varsinkin kasvojen reunoilla saattaa näkyä vääristymistä. Samoin videolla saattaa esiintyä yksittäisten kuvaruutujen välkkymistä tai epäluonnollisen näköistä valaistusta. (Sample 2020; Somers 2020.) Lisäksi videolla näkyvät kasvot saattavat näyttää sumeilta, eikä puhujan ääni välttämättä kuulosta vastaavan puhujan ulkoista olemusta (Johnson 2021). Synteettinen ääni saattaa myös kuulostaa epäluonnollisen tai metallisen kuuloiselta (Hallamaa 2021).

Reaaliaikaisen videoväärennöksen tunnistamiseen löytyy ainakin opinnäytetyön kirjoitushetkellä erillisiä keinoja. Videopuhelussa epäilyttävältä vaikuttavaa henkilöä voi esimerkiksi pyytää heiluttamaan kättään kasvojensa edessä, jolloin väärennöksen laatu todennäköisesti heikkenee hetkellisesti. Ennalta editoidussa syväväärennöksessä myös kasvojen edessä liikkuva käsi tai muu esine voidaan maskata siten, että väärennetyt kasvot näkyvät ainoastaan liikkuvan objektin takana. (Anderson 2022.)

Vielä tehokkaampi tapa reaaliaikaisen videoväärennöksen tunnistamiseen on henkilön pyytäminen kääntämään päätään 90 astetta sivulle, sillä sivuprofiilin väärentäminen on tekoälyalgoritmeille huomattavan haastavaa. Kasvonpiirteiden tunnistaminen ja automaattinen kohdistaminen eivät onnistu täydellisesti, kun puolet kasvoista ovat näkymättömissä. Tällöin tekoäly alkaa etsiä puuttuvia kasvonpiirteitä muualta kuvaruudusta, eikä kasvokuvasta tule tarkka. Tämän lisäksi sivuprofiilissa otetut kasvokuvat ovat yleisestikin harvinaisempia kuin edestäpäin otetut kuvat, jolloin myös tekoälyalgoritmeille on saatavilla vähemmän sivuprofiilikuvia, joista ne voisivat oppia sivuprofiilin väärentämistä. (Anderson 2022.)

Tekoälyn kehittyessä ihmiset eivät välttämättä voi enää tunnistaa uusia syväväärennöksiä (Schick 2020, 197). Tällöin ihmisten medialukutaidon tärkeys korostuu entisestään. Toimittaja ja tietokirjailija Johanna Vehkoo (2021) neuvoo tutustumaan epäilyttävän media-aineiston julkaisuyhteyteen. Hänen mukaansa kannattaa tarkastella, kuka aineiston on julkaissut ja mitä muuta he ovat

julkaisseet sekä pohtia julkaisijan motiiveja. Lisäksi hän neuvoo käyttämään Euroopan unionin rahoittamaa InVID-työkalua, jonka avulla voidaan tehdä käänteisiä kuvahakuja videon yksittäisistä kuvaruuduista. (Vehkoo 2021.) InVID-työkalu on ladattavissa internetselaimen laajennukseksi ilmaiseksi osoitteessa www.invid-project.eu. Työkalun avulla voidaan tehdä käänteisiä kuvahakuja Facebookissa, Instagramissa, YouTubessa, Twitterissä sekä Dailymotionissa julkaistuille videoille. (InVID Project 2022.) Myös tutkijoiden, yhteisöpalveluiden sekä journalistien tulee omalla työllään tukea mediakasvatusta ja yleistä medialukutaitoa. Tämä kuitenkin edellyttää, että kansa vielä luottaa näihin instituutioihin. (Somers 2020.) Kaikki voivat omalta osaltaan osallistua yleisen medialukutaidon ylläpitoon tukemalla luotettavia uutis- ja faktantarkistusorganisaatioita (Schick 2020, 192–195).

Syväväärennösten tunnistamiseen löytyy myös teknologisia ratkaisuja. Yhdysvaltain asevoimien tutkimusorganisaatio DARPA, Microsoft, Amazon ja Facebook (nykyisin Meta) ovat yhdessä tutkineet mahdollisuuksia käyttää tekoälyä syväväärennösten tunnistamiseen (Schick 2020, 195–196). Myös pienemmät teknologiayritykset pyrkivät tunnistamaan väärennöksiä tekoälyn avulla (Parkin 2019). Esimerkiksi varmennus- ja tunnistuspalvelu Sensity kertoo pystyvänsä tunnistamaan syväväärennökset yli 95 prosentin tarkkuudella (Cavalli 2021). Tällä hetkellä defensiiviset teknologiat ovat kuitenkin alakynnessä, minkä takia niiden täytyykin jatkaa kehittymistään väärennösteknologioiden rinnalla (Chesney & Citron 2019, 1787–1788). Toisaalta defensiivisten teknologioiden kehitys saattaa auttaa myös väärennöksiä kehittymään entisestään. Syväväärennösten tunnistaminen saattaakin tulevaisuudessa tulla mahdottomaksi myös tekoälyltä itseltään. (Schick 2020, 197.)

Toinen teknologinen vaihtoehto liittyy aidon media-aineiston todentamiseen. Uudet kamerat ja ääninauhurit voisivat sisällyttää tallentamaansa mediaan entistä tarkempaa metadataa sen alkuperäisestä tallennusajasta, -paikasta ja manipuloimattomuudesta (Nguyen ym. 2019, 9; Schick 2020, 197–198). Tutkijat ovat ehdottaneet myös sosiaalisen median yhteisöpalveluiden kuten Facebookin, Twitterin ja WhatsAppin ottavan suuremman vastuun hyväksymänsä sisällön tarkastamisesta ja suodattamisesta (Chesney & Citron 2019, 1817–1819).

Muun muassa Facebook (Bickert 2020), Twitter (2022), YouTube (Google 2022) ja TikTok (Pappas 2020) ovat sittemmin kieltäneet tarkoituksellisesti harhaanjohtavan, synteettisesti luodun tai manipuloidun median jakamisen omilla alustoillaan.

Uhkien kohtaamiseen tarvitaan myös laajempia, yhteiskunnallisia keinoja. Esimerkiksi Kiinassa ja yksittäisissä Yhdysvaltain osavaltioissa on jo säädetty syvävääreennösten levittämistä ja merkitsemistä sääntöleviä lakeja. Myös Euroopan parlamentti on huomioinut syvävääreennökset useissa esityksissään ja kannanotoissaan sekä ehdottanut, että syvävääreennösten tekijöille tulisi pakolliseksi merkitä jakamansa sisältö manipuloiduksi. Syvävääreennösten täyttä kieltoa ei tähän mennessä ole suunniteltu. Väärennösten täysi kieltäminen saattaisikin olla ongelmallista, sillä kaikki väärennökset eivät ole haitallisia. (Das ym. 2021, 37–68.) Täysi kieltäminen voisi myös hidastaa tekoälyn tutkimusta ja kehitystä (Chesney & Citron 2019, 1788–1789). Sekä yhteiskunnalliset että yhteisöpalvelujen tekemät ratkaisut vaativatkin haastavaa tasapainottelua turvallisuuden, yksityisyyden ja sananvapauden välillä (Schick 2020, 199–201).

6 Opinnäytetyön toiminnallisen osion esittely

Opinnäytetyön toiminnallinen osio tehtiin toimeksiantona Kansallisen audiovisuaalisen instituutin mediakasvatus- ja kuvaohjelmayksikölle. Toiminnallisen osion tavoitteena oli tuottaa tiivis mediakasvatusvideo, jonka avulla tuetaan lasten ja nuorten kriittistä medialukutaitoa. Videon avulla lapsia ja nuoria autetaan ymmärtämään, tunnistamaan ja käsittelemään syvävääreennöksiä misinformaatioon ja disinformaatioon liittyvänä ilmiönä mediayhteiskunnassa.

Mediakasvatusvideolle oli nähdäkseni oikea tarve, sillä syvävääreennökset ovat vielä tuore ja yleisesti verrattain tuntematon ilmiö. Nykyisessä tietoyhteiskunnassa mediasisältöjen kriittisen arvioinnin taidot ovat erityisen tärkeitä (Fagerlund ym. 2019, 58–59). Vaikka suomalaisnuorten medialukutaidon osaaminen on kansainvälisesti mitattuna hyvällä tasolla, vuonna 2018 tehdyn tutkimuksen

mukaan yli neljänneksellä suomalaisoppilaista oli kuitenkin heikko monilukutaidon taso. Tämä tarkoittaa, että heidän taitonsa tiedon löytämiseen, arvioimiseen ja hyödyntämiseen ovat hyvin puutteelliset, mikä saattaa vaarantaa heidät syrjäytymiselle opinnoista, työelämästä ja jopa laajemmastakin yhteiskunnasta. (Fagerlund ym. 2019, 17–19, 57.)

Opinnäytetyön toimeksiantaja toivoi mediakasvatusvideota virkeässä ja nuoria kiinnostavassa muodossa. Video sisältää myös leikkeitä itse tekemästäni videomanipulaatioesimerkistä. Video on suunnattu perusopetuksen 5.–9. luokan opiskelijoille ja pyrkii vastaamaan seuraaviin kysymyksiin:

- Mitä syvävääreennökset ovat?
- Miten manipulaatio tehdään?
- Miten manipulaation voi tunnistaa?
- Mitä haasteita ja vaikutuksia ilmiössä on?

Mediakasvatusvideo tuotettiin tukimateriaaliksi Kansallisen audiovisuaalisen instituutin ja Opetushallituksen yhteiseen Uudet lukutaidot -kehittämisohjelmaan, jonka tavoitteena on ”vahvistaa lasten ja nuorten medialukutaitoja, tieto- ja viestintäteknologista osaamista sekä ohjelmoinnin osaamista varhaiskasvatuksessa sekä esi- ja perusopetuksessa” (Uudet lukutaidot 2022a). Videon suunnittelussa hyödynnettiin kehittämisohjelmassa laadittuja perusopetuksen oppilaiden hyvää ja edistynyttä osaamista ilmaisevia medialukutaidon kuvauksia. Median tulkinnan ja arvioinnin kuvauksissa oppilaiden toivotaan kykenevän muun muassa pohtimaan mediasisältöjen tekijöiden motiiveja, arvioimaan mediasisältöjen luotettavuutta sekä tunnistamaan valheellisen ja harhaanjohtavan tiedon levittämiseen liittyviä mediailmiöitä (Uudet lukutaidot 2022b).

7 Mediakasvatusvideon tuotanto

7.1 Videon suunnittelu- ja käsikirjoitusvaiheet

Aloitin mediakasvatusvideon suunnittelemisen keskustelemalla videon sisällöstä ja toteutustavasta toimeksiantajani kanssa. Päädyimme keskustelussamme siihen, että videon tulisi selittää tiiviisti, mitä syväväärennökset ovat, mitä vaikutuksia niillä on, miten manipulaatio tehdään ja miten manipulaation voi tunnistaa. Sovimme, että video tulisi kestämään yhdestä kahteen minuuttia. Valitsin videon toteutustavaksi kaksiulotteisen liikegrafiikka-animaation, joka sisältäisi videomateriaalia myös omasta esimerkkivääärennöksestäni ja sen tekoprosessista. Videolla kuullaan aihetta selostavaa puhettani sekä taustamusiikkia.

Videon käsikirjoitusvaiheessa kirjoitin opinnäytetyöni tietoperustan avulla vastaukset edellisessä luvussa esitettyihin kysymyksiin. Pyrin myös muuttamaan vastausten kieliasua hieman yksinkertaisemmaksi, jotta niiden äänittäminen olisi luontevampaa. Samalla videon katselijat ymmärtäisivät videon sisällön helpommin.

Seuraavaksi suunnittelin selostuspuheeseeni alustavasti sopivia visualisointeja. Tulisin toteuttamaan osan visualisoinneista kaksiulotteisista vektorikuvakkeista animoituna liikegrafiikkana, ja osan hyödyntäen joko esimerkkimanipulaatiotani tai näiden yhdistelmää eli manipulaatiotani osana liikegrafiikkaa. Videomanipulaatiossani omat kasvoni muuttuvat Yhdysvaltain presidentti Joe Bidenin kasvoiksi. Valitsin videon toisiksi kasvoiksi juuri hänet, koska hän on hyvin laajasti tunnettu ja kasvonpiirteistään tunnistettava henkilö, jonka työnkuvaan kuuluvista julkisista esiintymisistä on saatavilla riittävästi vapaasti käytettävää videomateriaalia.

Kirjoitin käsikirjoitusdokumenttiini vierekkäisille palstoille videolla kuultavan selostuspuheen sekä sanallisen kuvauksen sitä tukevista visualisoinneista. En tehnyt erillistä kuvakäsikirjoitusta, sillä en vielä tässä vaiheessa tiennyt, miltä

videomanipulaationi tekoprosessi tulisi tarkalleen näyttämään. Myös liikegrafiikkana toteuttamani visualisoinnit tarkentuivat vielä tuotannon edetessä.

Lähetin käsikirjoitusdokumenttini toimeksiantajani kommentoitavaksi, ja he olivat siihen tyytyväisiä. He myös ehdottivat, että videolla voisi korostaa enemmänkin harhaanjohdetuksi tulemisen uhkaa ja mahdollisuutta aidon median kyseenalaistamiseen. Tämä oli mielestäni hyödyllinen korjaus, sillä alkuperäisessä käsikirjoituksessani mainitsin nämä vasta henkilökohtaisemman kiristetyksi tulemisen uhan jälkeen.

Suunnitteluvaiheen lopuksi minun oli selvitettävä videomanipulaation tekemiseen liittyviä tekijänoikeudellisia ja tietosuojalainsäädännöllisiä seikkoja, ennen kuin voisin aloittaa oman manipulaationi toteuttamisen. Yhdysvaltain tekijänoikeuslaissa säädetään, etteivät Yhdysvaltain hallituksen teokset ole tekijänoikeudella suojattuja (Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code / 2021, 105 §). Valkoisen talon tuottama videomateriaali on siis vapaasti uudelleenkäytettävissä. Kasvokuvia voidaan kuitenkin pitää yleisen tietosuoja-asetuksen alaisina henkilötietoina, jolloin niiden käsittely edellyttää lähtökohtaisesti laillista käsittelyperustetta. Lisäksi yleisessä tietosuoja-asetuksessa säädetään, että rekisterinpitäjän on informoitava rekisteröityä hänen henkilötietojensa käsittelystä (Euroopan parlamentin ja neuvoston asetus (EU) 2016/679, 12 §). Tämän takia presidentti Joe Bidenin henkilötietoja sisältävää materiaalia ei nähdä itse opinnäytetyöraportissa.

Olin kuitenkin aikeissa käsitellä kasvokuvia erityisesti mediakasvatuksellisen videon tuottamiseksi. Tämä voidaan tulkita journalistiseksi ilmaisuksi, mihin ei Tietosuojalain 1050/2018 (2018, 27 §) perusteella sovelleta tietosuoja-asetuksen artikloja esimerkiksi käsittelyperusteesta tai rekisteröidyn informointivelvollisuudesta. Tulkintaani mediakasvatusvideoni journalistisuudesta tukee myös Hallituksen esitys EU:n yleistä tietosuoja-asetusta täydentäväksi lainsäädännöksi (HE 9/2018 vp, 57), jossa journalistisesta ilmaisusta kirjoitetaan seuraavasti:

> Unionin tuomioistuimen oikeuskäytännössä henkilötietojen käsittelyksi journalistisessa tarkoituksessa on katsottu toiminta, jonka ainoana tarkoituksena on tietojen, mielipiteiden ja ajatusten ilmaiseminen yleisölle.

> Unionin tuomioistuin on korostanut, että merkitystä sitä vastoin ei ole sillä, liittyykö toimintaan voitontavoittelu, eikä liioin tavalla, jolla tiedot siirretään. Lisäksi oikeuskäytännössä on painotettu, että henkilötietojen käsittelyä koskevat poikkeukset koskevat joukkotiedotusyritysten lisäksi kaikkia journalismia harjoittavia henkilöitä. (Hallituksen esitys HE 9/2018 vp, 57.)

Tiivistettynä edellisessä sitaatissa siis vahvistetaan, että myös yksityishenkilön audiovisuaalisesti yleisölle välittämää informaatiota pidetään Unionin tuomioistuimen oikeuskäytännössä journalismina. Tämän perusteella voin käsitellä henkilötietoja mediakasvatusvideollani hieman vapaammin. Journalistiliiton lakimies Hannu Hallamaa (2019) kuitenkin huomauttaa, ettei journalistinen poikkeus vapauta velvoitteesta huolehtia tietoturvasta. Henkilötietoja on siis joka tapauksessa käsiteltävä huolellisesti.

7.2 Oman syväväärennösesimerkkini tekoprosessi

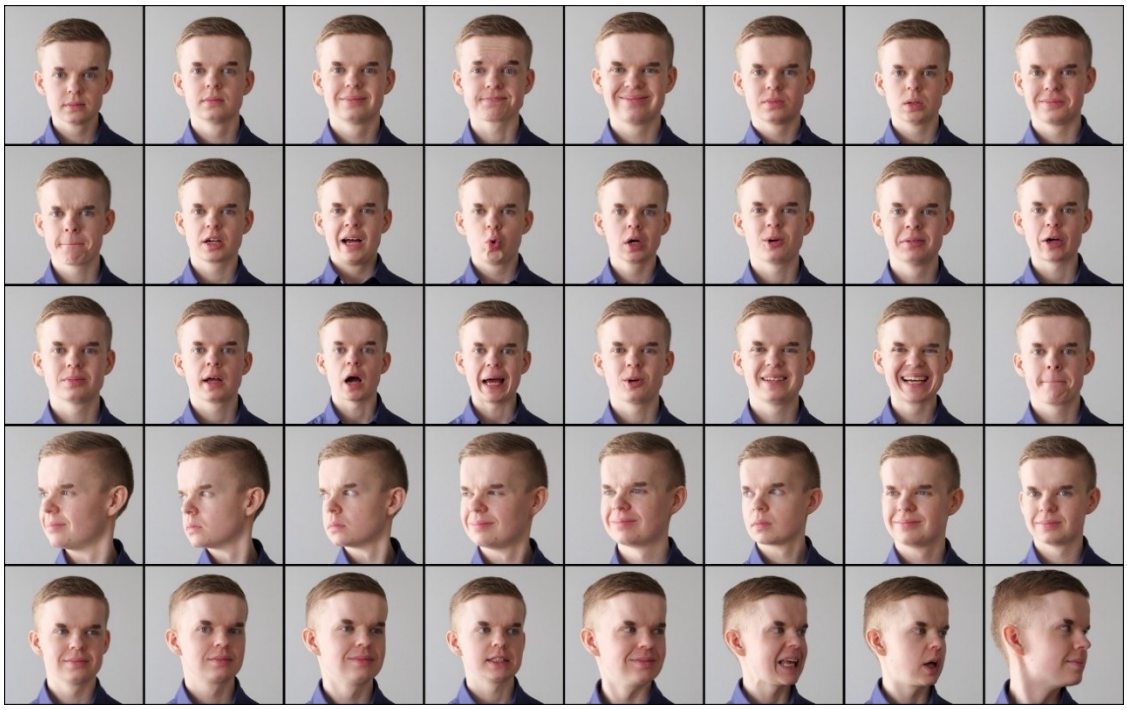
Suunnitteluvaiheen jälkeen latasin videotallenteita Joe Bidenin julkisista esiintymisistä Yhdysvaltain presidenttiyttä seuraavan ja taltioivan, Virginian yliopiston yhteydessä toimivan Miller Centerin verkkosivustolta. Videotallenteiden lähteeksi on merkitty Valkoinen talo, mutta varmistin videoiden tekijänoikeudettomuuden ja käyttöluvan vielä Miller Centerin verkkosivustolta löytyvän yhteydenottolomakkeen kautta, ja sain tähän vahvistuksen sähköpostitse (Greco 2022).

Leikkasin videotallenteista kymmenen minuutin pituisen koosteen, jossa Joe Bidenin kasvoja on kuvattu mahdollisimman monipuolisesti. Videomateriaalin kuvataajuus oli 30 kuvaruutua sekunnissa. Siirsin leikkaamani koosteen Faceswap-ohjelmistoon, jolla irrotin videomateriaalista automaattisesti joka kahdeksannen kuvaruudun. Valitsin tämän, jotta Bidenin ilme ehtisi muuttua perättäisten kasvokuvien välillä, ja jottei minulle kertyisi suurta määrää lähes identtisiä kasvokuvia.

Faceswap-ohjelmisto tarjoaa kasvojen tunnistamiseen, paikantamiseen ja rajaamiseen useita asetuksia, joista käytin ohjelmiston oletusasetuksia, sillä

pyrin omalla syväväärennökselläni myös osoittamaan, ettei väärennöksen tekemiseen vaadita erityisen laajaa teknistä perehtymistä. Oletusasetuksia kuvattiin ohjelmiston sisäisissä vihjeteksteissä parhaiksi, joskin näytönohjaimen tehoa vaativiksi vaihtoehdoiksi. Faceswap tunnisti, paikansi, rajasi ja irrotti videomateriaalista 2248 kasvokuvaa sekä 43 virheellistä kuvaa, joissa ei näkynyt ollenkaan kasvoja, ja tallensi kasvojen sijaintitiedot erilliseen alignments-tiedostoon. Poistin virheelliset kuvat manuaalisesti, minkä jälkeen poistin myös poistettujen kuvien sijaintitiedot Faceswapin Alignments-työkalulla, sillä niistä voisi olla haittaa tekoälyn kouluttamisessa (Torzdf 2022a).

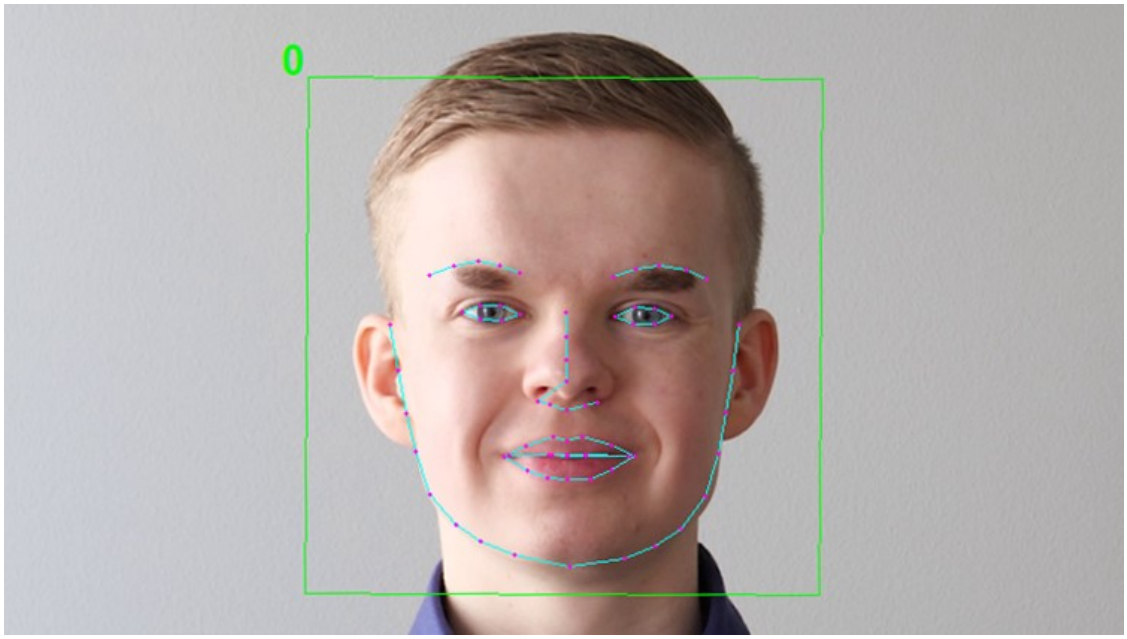
Seuraavaksi kuvasin noin yhdeksän ja puoli minuuttia videomateriaalia omista kasvoistani. Koska Joe Biden ei videoesiintymisissään pitänyt silmälaseja, kuvasin myös omia kavojani ilman silmälaseja. Silmälasit saattaisivat myös tuottaa tekoälylle ongelmia, kuten opinnäytetyön tietoperustassa ilmeni. Asetin Panasonicin LUMIX DC-G9 -järjestelmäkamerani jalustalle ja kuvasin kasvojani monipuolisista suunnista tehden samalla vaihtelevia ilmeitä (kuva 1).



Kuva 1. Kasvoni kuvattuna eri suunnista vaihtelevilla ilmeillä.

Kuvasin videon kuvataajuudella 25 ruutua sekunnissa, ja irrotin jälleen joka kahdeksannen kuvaruudun Faceswapilla. Tästä kertyi 1772 kasvokuvaa

itsestäni, joista yksi oli ylösalaisin. Poistin virheellisen kuvan ja sen sijaintitiedot, jotta ne eivät aiheuttaisi ongelmia tekoälyn koulutusvaiheessa. Kasvokuvien automaattinen tunnistaminen, paikantaminen, rajaus ja irrottaminen oli nopea ja vaivaton työvaihe, johon kului vain muutama minuutti. Lopulliset kasvokuvat kannatti kuitenkin tarkistaa itse, sillä minun tapauksessani molempien henkilöiden kasvokuvien seassa oli virheellisiä kuvia. Faceswapin paikantamat kasvonpiirteeni on kuvattu kuvassa 2.

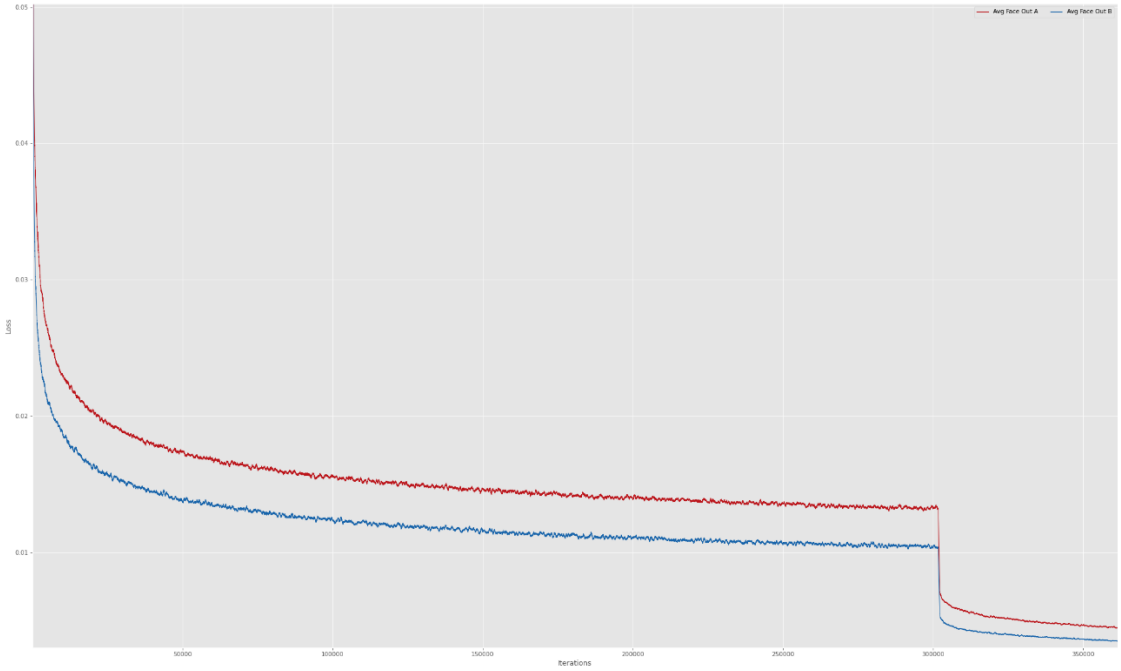


Kuva 2. Faceswap-ohjelmiston tunnistamat ja paikantamat kasvonpiirteeni.

Kasvokuvien irrottamisen jälkeen pystyin aloittamaan tekoälymallin kouluttamisen. Faceswap tarjoaa tähän useita eri tekoälymalleja omine asetuksineen. Valitsin oletuksena olleen Original-mallin sijasta Dfaker-mallin, sillä Original-mallilla voi luoda syväväärennettyjä kasvokuvia vain 64 pikselin levyisinä ja korkuisina. Dfaker-mallissa kasvokuvien maksimikoko sen sijaan on leveydeltään ja korkeudeltaan 256 pikseliä. Valitsin kuitenkin oman väärennökseni kasvokuvien kooksi 128 x 128 pikseliä, sillä Valkoisen talon videotallenteiden resoluutio oli vain 1280 x 720 pikseliä, minkä vuoksi myös Joe Bidenin kasvokuvat jäivät melko pienikokoisiksi. Lisäksi väärennöksen haluttu koko vaikuttaa olennaisesti koulutusprosessin kestoon. Kasvokuvien leveyden ja korkeuden kaksinkertaistuessa tekoälyn opiskeleman alueen pikselimäärä ja täten myös koulutusaika nelinkertaistuu (Torzdf 2022b).

Dfaker-mallin kuvataan Faceswapin käyttöohjeissa kykenevän laadukkaisiin lopputuloksiin ilman oletusasetusten muuttamista (Torzdf 2022b), joten se vaikutti sopivan omaan käyttötarkoitukseeni varsin hyvin. Pidin tekoälylle syötettävien kuvien kertaerän oletusasetusten mukaisena 16 kappaleena ja aloitin tekoälyn kouluttamisen. Kytkin myös Faceswapin timelapse-asetuksen päälle, jotta ohjelmisto tallentaisi kuvia koulutusprosessin kehityksestä. Voisin myöhemmin hyödyntää näitä timelapse-kuvia mediakasvatusvideollani. Koulutin neuroverkkoa yhteensä 24 tunnin ajan noin kahden viikon ajanjaksolla, minkä aikana Faceswap suoritti 360 000 iteraatiota. Tämä tarkoittaa, että tekoälymalli näki ja uudelleenloi 360 000 kuudentoista kasvokuvan kertaerää eli yli 5,7 miljoonaa kasvokuvaa (Torzdf 2022b).

Faceswap kuvaa neuroverkkomallin oppimista loss-arvolla, joka kuvaa sitä, kuinka oikeita kasvokuvia vastaavina malli pitää uudelleenluomiaan kasvokuvia. Mitä lähemmäs nollaa loss-arvo laskee, sitä aidomman oloisina malli pitää luomiaan kasvokuvia. (Torzdf 2022b.) Kuvio 3 kuvaa, kuinka loss-arvo laski koulutusprosessini aikana noin 0,05:stä alle 0,005:een. Arvot eivät itsessään ole erityisen merkityksellisiä tai vertailukelpoisia muihin tekoälyn koulutusprosesseihin, mutta ne kuitenkin kuvaavat tämän yksittäisen koulutusprosessin aikana tapahtunutta kehitystä. 300 000 iteraation eli noin 20 tunnin jälkeen tapahtunut loss-arvon äkillinen laskeminen eli kasvokuvien laadun paraneminen johtuu siitä, että kytkin No Warp -asetuksen päälle Faceswapissa. Kuvan vääristämisen mainitaan Faceswapin ohjeissa olevan erittäin tärkeää neuroverkon oppimiselle, mutta koulutusprosessin loppua kohden kuvien vääristämättä jättämisen kerrotaan mahdollisesti auttavan tuomaan esiin kasvojen yksityiskohtia (Torzdf 2022b).



Kuvio 3. Loss-arvon kehitys tekoälyn koulutusprosessini aikana.

Tekoälymallin koulutusprosessilla ei ole selkeää loppua, vaan neuroverkon kouluttaminen on lopetettava itse, kun väärennöksen laatu alkaa vaikuttaa riittävän hyvältä (Torzdf 2022b). Olisin voinut jatkaa tekoälyn kouluttamista vielä pidempäänkin, mutta aloin olla jo tyytyväinen Faceswap-ohjelmiston esikatseluikkunassa näkyvään lopputulokseen (kuva 3). Lisäksi pidin 24:ää tuntia rajana, jota pidempään en usko aiheeseen perehtymättömien henkilöiden viitsivän sitoutua mahdollisissa omissa väärennöksissään. Oletin itsekin, että väärennökseni olisi valmistunut tätä nopeammin. Omassa tietokoneessani on myös tehokas pelaamiseen ja luovaan työhön suunniteltu NVIDIA GeForce RTX 2080 Ti -näytönohjain, mikä varmasti nopeutti tekoälyn koulutusprosessia. Hitaammalla tietokoneella ja väärennösohjelmiston eri asetuksilla koulutusprosessi olisi siis voinut kestää vielä huomattavasti pidempäänkin.

Kuva 3. Alkuperäiset kasvokuvani sekä tekoälyn uudelleenluomat kuvat kasvoistani koulutusprosessin alussa ja lopussa.

Viimeinen vaihe syväväärennökseni tekoprosessissa oli Joe Bidenin kasvokuvien siirtäminen omien kasvojeni tilalle alkuperäiseen videoon. Koska kasvokuvien irrottamisen yhteydessä luotuun alignments-tiedostoon oli tallennettu kasvojeni sijainti vain videon joka kahdeksannessa kuvaruudussa, täytyi nyt tehdä uusi alignments-tiedosto, joka sisältäisi kasvojeni sijaintitiedot videon jokaisen kuvaruudun osalta. Muuten Bidenin kasvot vilkahtaisivat lopullisella videolla vain joka kahdeksannessa kuvaruudussa. Uuden alignments-tiedoston luominen Faceswapissa kesti tällä kertaa noin 30 minuuttia.

Tämän jälkeen säädin vielä kasvojen väri- ja yhdistämisasetuksia Faceswapin Preview-työkalulla, jotta Bidenin kasvot sopisivat alkuperäiseen videooni paremmin. Lopuksi eksportoin valmiin syväväärennöksen 1280 x 720 -resoluutiossa. Valmis syväväärennös ei ole laadultaan täydellinen, mutta kuitenkin riittävän hyvä mediakasvatusvideoon sisällytettäväksi. Paikoitellen väärennös näyttää mielestäni yllättävänkin hyvältä. Sivuprofiili on

väärennöksessäni kuitenkin hyvin epäonnistunut, kuten oli arvattavissa opinnäytetyön tietoperustan perusteella.

7.3 Mediakasvatusvideon tuotanto- ja editointivaiheet

Mediakasvatusvideoni tuotantovaihe alkoi videolla kuultavan selostuspuheeni äänittämisellä. Suoritin äänittämisen äänieristetyssä työskentelykopissa kannettavalla Zoom H2N -stereotallentimella. Tallennin oli helppokäyttöinen, mutta puheeni äänittäminen ei ollut itselleni luontaisen oloinen työvaihe. Äänitin puheeni kuitenkin useaan kertaan, jotta sain myöhemmin valikoitua ja editoitua onnistuneimmat äänitallenteeni yhteen Adobe Premiere Pro -videonmuokkausohjelmistolla. Latasin mediakasvatusvideolla kuultavan musiikin ja kaksi äänitehostetta YouTuben audiokirjastosta sekä yhden erillisen äänitehosteen 99sounds.org-verkkosivustolta. Molemmat sivustot sallivat äänitiedostojen hyödyntämisen sekä henkilökohtaisiin että kaupallisiin tarkoituksiin ilman erillistä mainintaa äänten alkuperästä.

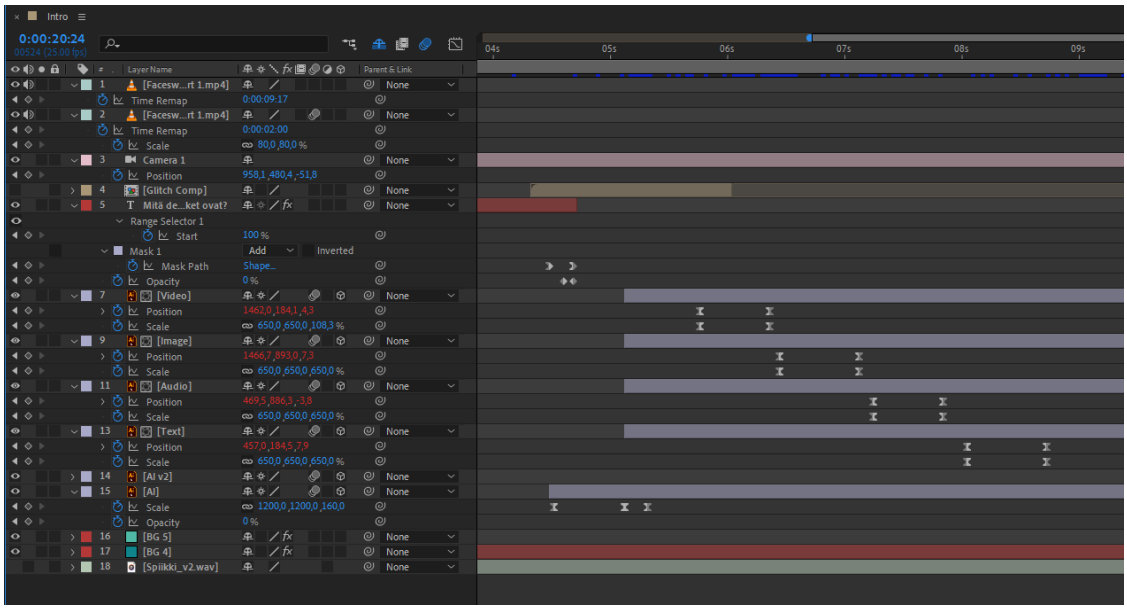
Seuraavaksi aloin kokoamaan mediakasvatusvideon tuottamiseen tarvitsemiani erinäisiä visuaalisia materiaaleja. Koostin Faceswap-ohjelmiston tallentamista timelapse-kuvista lyhyen videoklipin, jossa nähdään keinotekoisten kasvokuvien kehitys tekoälyn koulutusprosessin aikana. Lisäksi leikkasin luomastani syvävääreennöksestä lyhyempiä, mediakasvatusvideoon soveltuvia leikkeitä. Editoin timelapse- ja syvävääreennösleikkeet Adobe Premiere Pro:lla. Koostin myös Faceswapin aiemmin irrottamista kasvokuvista erikokoisia kuvakoosteita Adobe InDesign -taitto-ohjelmistolla. Tämän jälkeen latasin vapaasti käytettäviä ja muokattavia kaksiulotteisia vektorikuvakkeita iconmonstr.com-verkkosivustolta sekä muokkasin kuvakkeita Adobe Illustrator-vektorigrafiikkaohjelmistolla. Viimeiseksi poimin Uudet lukutaidot -verkkosivustolla käytettyjen värien RGB-värikoodit Adobe Color -verkkopalvelun avulla.

Kun olin kerännyt tarvitsemani materiaalit, pääsin aloittamaan materiaalien animoinnin Adobe After Effects -liikegrafiikkaohjelmistolla. Toin Premiere Pro:lla editoimani ääniraidan After Effects -projektiin ja aloin animoimaan

kaksiulotteisia muotoja, tekstiä, kuvakkeita, kuvakoosteita sekä videoklippejä selostuspuheeni tahtiin. Olin käyttänyt After Effects -ohjelmistoa aiemmin omien liikegrafiikkaharjoitusteni tekemiseen, joten erilaisten objektien animointi oli minulle ennestään tuttua. Mediakasvatusvideoni oli kuitenkin huomattavasti aiempia harjoitusprojektejani pidempi.

Kaksi minuuttia kestävän mediakasvatusvideon tekeminen oli melko työlästä, sillä video koostuu lukuisista samanaikaisista, limittäisistä ja peräkkäisistä animaatioista. Animoin useiden eri objektien monipuolisia ominaisuuksia, kuten kokoa, sijaintia, läpinäkyvyyttä, kiertoa sekä maskia. Animoin objektien ominaisuuksia After Effects -ohjelmiston avainruuduilla, joilla määritetään ominaisuuden arvo senhetkisessä kuvaruudussa. Kun saman ominaisuuden kahteen avainruutuun määritetään eri arvo, syntyy näiden kuvaruutujen välille animaatio, jossa arvo muuttuu (Adobe 2022). Avainruutujen lisäksi hyödynsin mediakasvatusvideon animoinnissa maskeja, virtuaalisia kameroita sekä sisäkkäisiä kompositioita, joiden avulla sain animoitua useampia objekteja kerralla.

Jaoin mediakasvatusvideon neljään After Effects -kompositioon, jotka kaikki sisälsivät myös pienempiä, sisäkkäisiä kompositioita. Pienemmissä osissa koen saaneeni organisoitua animaatioprojektin sujuvammin. Kuva 4 havainnollistaa, kuinka monia tasoja ja avainruutuja näennäisen yksinkertaisenkin animaation tekemiseen vaaditaan. Kuvan aikajana käsittää noin viiden sekunnin mittaisen osion mediakasvatusvideon alusta.



Kuva 4. Mediakasvatusvideoni tekemiseen käytettyjä tasoja ja avainruutuja Adobe After Effects -aikajanalla.

Lopuksi editoin After Effects -kompositiot, selostuspuheeni, taustamusiikin ja äänitehosteet yhteen Adobe Premiere Pro:lla. Hienosäädin vielä animaatioiden ajoituksia, kun olin saanut tarkasteltua videota valmiimpana kokonaisuutena. Adoben Creative Cloud -ohjelmistojen Dynamic Link -ominaisuus oli tässä hyödyllinen, sillä sain muutettua animaatioita After Effects:ssä, jolloin ne päivittyivät suoraan myös lopulliseen Premiere Pro -projektitiedostooni ilman animaatioiden eksportointia jokaisen muutoksen jälkeen. Tietyt animaation osat tosin toistuivat tietokoneellani hyvin hitaasti ilman niiden erillistä renderöintiä, minkä vuoksi videon kannatti tarkistaa lopullisesti vasta eksportoinnin jälkeen.

Lähetin mediakasvatusvideon ensimmäisen version toimeksiantajani kommentoitavaksi. Sain videosta kehuja, mutta myös pyynnön lisätä videoon tekstinostot vinkeistä, joiden avulla videomanipulaation voi tunnistaa. Lisäksi toimeksiantajani pyysi tarkistamaan videolla näkyvien tekstien saavutettavuuden. Saavutettavuuden tarkistaminen oli hyödyllinen kommentti, jota en ollut etukäteen osannut ajatella. Tarkistin videolla nähtävien tekstien kontrastin Adobe Color -verkkopalveluun kuuluvalla kontrastin tarkistustoiminnolla ja muutin sen perusteella videoni tummia tekstejä vielä aiempaa tummemmiksi. Lisäsin myös tekstinostot videoni manipulaation tunnistamista käsittelevään osioon ja tein vielä viimeisiä hienosäätöjä animaatioihin. Tämän jälkeen mediakasvatusvideoni oli valmis (liite 1).

8 Tulokset

Opinnäytetyön tuloksena syntyi syväväärennöksiä käsittelevä kahden minuutin pituinen mediakasvatusvideo sekä kirjallinen tietoperusta, jotka yhdessä antavat lukijalle monipuolisen perustietämyksen syväväärennöksistä. Tietoperustan ja mediakasvatusvideon avulla lukija toivottavasti kykenee arvioimaan kohtaamiensa mediasisältöjen luotettavuutta sekä tunnistamaan valheellisen ja harhaanjohtavan tiedon levittämiseen liittyviä mediailmiöitä.

Mediakasvatusvideolla nähtävä syväväärennös ei ole erityisen laadukas tai vakuuttava, mutta se kuitenkin antaa katsojalle käytännön esimerkin siitä, mitä uudenlaisella teknologialla pystytään tekemään. Opinnäytetyön toimeksiantaja on ilmaissut olevansa tyytyväinen valmiiseen videoon.

Valmis mediakasvatusvideo julkaistiin osana Uudet lukutaidot -kehittämisohjelman tukimateriaaleja Asiaa mediakasvatuksesta -YouTube-kanavalla ja Mediataitokoulun verkkosivustolla. Videon sisältävä Mediataitokoulun verkkosivujulkaisu on linkitetty myös Uudet lukutaidot -verkkosivuston Polkuja uusiin mediailmiöihin -osioon.

9 Pohdinta

Opinnäytetyön valmis tietoperusta on mielestäni kattava kokonaisuus, joka keskittyy media-alan koulutuksen kannalta olennaisiin asioihin. Aihetta käsitellään mediayhteiskunnassa nähtävänä ilmiönä eikä niinkään teknologian kannalta, vaikka tekoälyn toimintaa selostetaankin opinnäytetyössä pintapuolisesti. En kokenut tärkeäksi perehtyä tekoälyn toimintaan tämän yksityiskohtaisemmin, sillä se ei suoranaisesti liity media-alan opintoihini. Oma ymmärrykseni koneoppimisen teknisestä ulottuvuudesta on myös hyvin rajallista. Lisäksi tekoälysovellukset kehittyvät nopeasti, jolloin tekniset yksityiskohdat eivät välttämättä ole relevantteja pitkään.

Myös syväväärennösteknologia kehittyy ja yleistyy vauhdilla, minkä takia aihetta kannattaa seurata jatkossakin. Aihetta käsittelevissä jatkotutkimuksissa olisi mahdollista perehtyä tarkemmin ilmiön teknologiseen kehitykseen sekä mediassa leviäviin konkreettisiin väärennöksiin. Myös äänen manipulointia olisi mahdollista tutkia tarkemmin, varsinkin kun äänen väärennösteknologia tulee aiempaa yleisemmin saataville.

Tuottamani syväväärennös ei itsessään ole erityisen vakuuttava tai lopulta edes tunnistettavasti presidentti Joe Bidenin kasvoja kuvaava, mutta mielestäni se toimii kuitenkin hyödyllisenä esimerkkinä osana mediakasvatusvideotani. Videolla näkyy selvästi, etteivät manipuloidut kasvot ole omani. Laadukkaamman väärennöksen luomiseksi kannattaisi luultavasti itse kuvata molempia henkilöitä esittävät videot, valita lopulliselle manipulaatiolle suurempi kuvatarkkuus ja antaa tekoälyn kouluttamisprosessin jatkua huomattavasti pidempään.

Opin omasta syväväärennösprosessistani paljon, ja olen tyytyväinen siihen, että sain tehtyä oman manipulaationi ilman tämän suurempaa perehtymistä väärennösohjelmiston tai tekoälyalgoritmien teknisiin toimintaperiaatteisiin. Mielestäni toteuttamani syväväärennös toimii myös hyvänä esimerkkinä siitä, kuinka käytännössä kuka vain pystyisi tekemään oman syväväärennöksensä. Laadukkaampaan lopputulokseen olisi kuitenkin vaadittu hieman enemmän aikaa ja vaivannäköä. Tekoälyn koulutusprosessi vei niin kauan aikaa, etten itse usko tekeväni enää uutta esimerkkiväärennöstä ainakaan nykyisellä väärennösteknologialla.

Olen tyytyväinen luomaani mediakasvatusvideoon, ja myös opinnäytetyön toimeksiantaja on kehunut videota. En yleensäkään pidä oman ääneni jälkikäteen kuuntelemista erityisen mielekkäänä, mutta videon selostuspuhe kuulostaa omaan korvaani varsinkin YouTubesta kuunneltuna hieman terävältä. Puheen olisi luultavasti voinut äänittää ja käsitellä laadukkaammin. Myös mediakasvatusvideon kuvitus olisi tietyiltä osin voinut olla hieman kiinnostavampaa, mutta kokonaisuutena olen videoon oikein tyytyväinen.

Opinnäytetyöprosessini oli melko pitkä ja katkonainen, mikä ilmenee myös opinnäytetyön lähdeluettelon viittauspäivämääristä. Opinnäytetyön tekeminen kesti kokonaisuudessaan lähes kaksi vuotta. Olin suuren osan tästä ajasta joko työharjoittelussa tai töissä, jolloin en jaksanut tai ehtinyt edistää opinnäytetyötäni. Työ eteni kuitenkin hyvin aina kun aloitin sen työstämisen uudelleen. Opinnäytetyön toiminnallinen osa valmistui muutamassa kuukaudessa keväällä 2022, ja kommunikaatio työn toimeksiantajan kanssa sujui hyvin. En kohdannut opinnäytetyötä tehdessäni varsinaisia haasteita, ja opin opinnäytetyöprosessistani paljon. Prosessin kestosta huolimatta olen opinnäytetyöhöni varsin tyytyväinen.

Lähteet

Adobe. 2022. Keyframe animation for beginners. https://www.adobe.com/creativecloud/video/discover/keyframing.html. 25.9.2022.

Ajder, H., Patrini, G., Cavalli, F. & Cullen, L. 2019. The State of Deepfakes: Landscape, Threats, and Impact. Sensity. https://sensity.ai/reports/. 19.11.2020.

Anderson, M. 2021. Real-Time DeepFake Streaming With DeepFaceLive. Unite.AI. https://www.unite.ai/real-time-deepfake-streaming-with-deepfacelive/. 19.9.2022.

Anderson, M. 2022. To Uncover a Deepfake Video Call, Ask the Caller to Turn Sideways. Metaphysic. https://metaphysic.ai/to-uncover-a-deepfake-video-call-ask-the-caller-to-turn-sideways/. 13.9.2022.

Bickert, M. 2020. Enforcing Against Manipulated Media. Meta. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/. 24.2.2022.

Brandom, R. 2019. Deepfake propaganda is not a real problem. The Verge. 5.3.2019. https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax. 15.2.2021.

BuzzFeedVideo. 2018. You Won't Believe What Obama Says In This Video! https://youtu.be/cQ54GDm1eL0. YouTube-video. 29.4.2021.

Cavalli, F. 2021. How to detect a deepfake online. Sensity. https://sensity.ai/how-to-detect-a-deepfake/. 15.4.2021.

Channel 4. 2020. Deepfake Queen: 2020 Alternative Christmas Message. https://youtu.be/IvY-Abd2FfM. YouTube-video. 29.4.2021.

Chesney, B. & Citron, D. 2019. Deep Fakes: A Looming Challenge for Privacy. California Law Review, 1753–1820. https://lawcat.berkeley.edu/record/1136469. 20.11.2020.

Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code / 2021.

Das, D., Fatun, M., Gerritsen, J., Jahnel, J., Karaboga, M., Kool, L., Nierling, L., van Boheemen, P. & van Huijstee, M. 2021. Tackling deepfakes in European policy. Panel for the Future of Science and Technology (STOA). https://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2021)690039. 23.2.2022.

Deepfakes. 2021. FaceSwap has ethical uses. https://github.com/deepfakes/faceswap/#faceswap-has-ethical-uses. 8.2.2022.

Eckart, K. 2021. A growing problem of 'deepfake geography': How AI falsifies satellite images. University of Washington. 21.4.2021. https://www.washington.edu/news/2021/04/21/a-growing-problem-of-deepfake-geography-how-ai-falsifies-satellite-images/. 28.4.2021.

Eronen, R., Haapanen, M., Heinonen, T., Joki, L., Klemettinen, R., Leskelä, H., Paajanen, I. & Pyhälahti, M. 2019. Sanapoimintoja vuodelta 2019. Kotimaisten kielten keskus. https://www.kotus.fi/sanakirjat/kielitoimiston_sanakirja/uudet_sanat/vuoden_sanapoiminnot/sanapoimintoja_2019. 16.2.2021.

Euroopan parlamentin ja neuvoston asetus (EU) 2016/679.

Faceswap. 2022. Welcome - Faceswap. https://faceswap.dev/. 8.2.2022.

Fagerlund, J., Leino, K., Niilo-Rämä, M., Puhakka, E., Rikala, J. & Sirén, M. 2019. Digiloikasta digitaitoihin. Kansainvälinen monilukutaidon ja ohjelmoinnillisen ajattelun tutkimus (ICILS 2018). Koulutuksen tutkimuslaitos. https://ktl.jyu.fi/fi/julkaisut/julkaisuluettelo-1/julkaisujen-sivut/2019/icils-2018-raportti.pdf. 23.2.2022.

Flynn, J. 2021. Helping actor Val Kilmer reclaim his voice. Sonantic. https://www.sonantic.io/blog/helping-actor-val-kilmer-reclaim-his-voice. 7.9.2022.

Fowler, G. 2021. Anyone with an iPhone can now make deepfakes. We aren't ready for what happens next. The Washington Post. 25.3.2021. https://www.washingtonpost.com/technology/2021/03/25/deepfake-video-apps/. 15.4.2021.

Gerhart, A. 2021. Election results under attack: Here are the facts. The Washington Post. 11.3.2020. https://www.washingtonpost.com/elections/interactive/2020/election-integrity/. 15.2.2021.

Giardina, C. 2022. SAG-AFTRA: Deepfakes "Pose a Potential Threat to Performers' Livelihoods". The Hollywood Reporter. 21.7.2022. https://www.hollywoodreporter.com/business/digital/sag-aftra-deepfakes-performers-threat-1235182933/. 7.9.2022.

Google Trends. 2022a. deepfake - Tutki - Google Trends. https://trends.google.com/trends/explore?date=today%205-y&q=deepfake. 17.9.2022.

Google Trends. 2022b. deepfake - Tutki - Google Trends. https://trends.google.com/trends/explore?date=2021-02-14%202021-03-31&q=deepfake. 18.9.2022.

Google. 2022. Misinformation policies – YouTube Help. https://support.google.com/youtube/answer/10834785. 21.3.2022.

Greco, M. 2022. [Send us a message]. joonas.jutila@edu.karelia.fi. 23.2.2022.

Hallamaa, H. 2018. Lakimies Hannu Hallamaa: Tietosuoja-asetus vaikuttaa journalistien työhön vain vähän. Journalistiliitto. https://journalistiliitto.fi/fi/lakimies-hannu-hallamaa-tietosuoja-asetus-vaikuttaa-journalistien-tyohon-vain-vahan/. 8.4.2022.

Hallamaa, T. 2018. Neuroverkko katsoi kuvia julkkiksista ja alkoi luoda kasvoja, joita ei ole olemassa – Tunnustammeko koneen älyn vasta, kun se kykenee huijaamaan meitä? Yle. 15.3.2018. https://yle.fi/uutiset/3-10115902. 22.2.2022.

Hallamaa, T. 2019. Tämän jutun jälkeen katsot liikkuvaa kuvaa uusin silmin: Yle teki deepfake-videon, jolla Sauli Niinistö haaveilee kolmannesta kaudesta. Yle. 6.9.2019. https://yle.fi/uutiset/3-10955498. 19.11.2020.

Hallamaa, T. 2021. Väärensimme Jenni Poikeluksen äänen ja otimme hänelle lopputilin: "Pitäkää tunkkinne" – samalla selvisi, mistä deepfake-ilmiössä on kyse. Yle. 8.9.2021. https://yle.fi/uutiset/3-12044625. 16.2.2021.

Hallituksen esitys HE 9/2018 vp.

Hao, K. 2019. The biggest threat of deepfakes isn't the deepfakes themselves. MIT Technology Review. 10.10.2019. https://www.technologyreview.com/2019/10/10/132667/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/. 19.11.2020.

Hao, K. 2020. Inside the strange new world of being a deepfake actor. MIT Technology Review. 9.10.2020. https://www.technologyreview.com/2020/10/09/1009850/ai-deepfake-acting/. 19.11.2020.

Holroyd, M. & Olorunselu, F. 2022. Deepfake Zelenskyy surrender video is the 'first intentionally used' in Ukraine war. Euronews. 16.3.2022. https://www.euronews.com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war. 17.3.2022.

Internet Crime Complaint Center (IC3). 2022. Deepfakes and Stolen PII Utilized to Apply for Remote Work Positions. https://www.ic3.gov/Media/Y2022/PSA220628. 19.9.2022.

InVID Project. 2022. InVID Verification Application. https://www.invid-project.eu/tools-and-services/invid-verification-plugin/. 17.10.2022.

Jantunen, S. 2016. Hallintovaliokunnan kuuleminen 30.9.2016 klo 11:15. Puolustusvoimien tutkimuslaitos. https://www.eduskunta.fi/FI/vaski/JulkaisuMetatieto/Documents/EDK-2016-AK-76663.pdf. 28.4.2021.

Johnson, D. 2021. What is a deepfake? Everything you need to know about the AI-powered fake media. Business Insider. 22.1.2021. https://www.businessinsider.com/what-is-deepfake?. 28.2.2022.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M. & Theobalt, C. 2018. Deep video portraits. ACM Trans. Graph. 37 (4). Article 163. https://doi.org/10.1145/3197517.3201283. 19.11.2020.

Korshunov, P. & Marcel, S. 2019. Vulnerability assessment and detection of Deepfake videos. 2019 International Conference on Biometrics (ICB), 1–6. https://ieeexplore.ieee.org/document/8987375. 20.11.2020.

Nguyen, T., Nguyen, C., Nguyen, T., Nguyen, D., Nahavandi, S., Pham, Q. & Huynh-The, T. 2019. Deep Learning for Deepfakes Creation and Detection: A Survey. https://www.researchgate.net/publication/336055871_Deep_Learning_for_Deepfakes_Creation_and_Detection_A_Survey. 20.11.2020.

Pappas, V. 2020. Combating misinformation and election interference on TikTok. TikTok. https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok. 24.2.2022.

Paris, B. & Donovan, J. 2019. Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence. Data & Society. https://datasociety.net/library/deepfakes-and-cheap-fakes/. 19.11.2020.

Parkin, S. 2019. The rise of the deepfake and the threat to democracy. The Guardian. 22.6.2019. https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy. 23.11.2020.

Rosner, H. 2021. The Ethics of a Deepfake Anthony Bourdain Voice in “Roadrunner”. The New Yorker. 17.7.2021. https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice. 22.2.2022.

Sample, I. 2020. What are deepfakes – and how can you spot them? The Guardian. 13.1.2020. https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them. 19.11.2020.

Schick, N. 2020. Deep Fakes and the Infocalypse – What You Urgently Need to Know. Lontoo: Monoray.

Sensity. 2022. Deepfakes vs Biometric KYC Verification. 13.9.2022.

Siltanen, S. 2018. Algoritmi toimii kuin anopin kakkuresepti – Miksi se sitten pelottaa niin paljon? Yle. 8.6.2018. https://yle.fi/aihe/artikkeli/2018/06/08/algoritmi-toimii-kuin-anopin-kakkuresepti-miksi-se-sitten-pelottaa-niin-paljon. 27.9.2022.

Somers, M. 2020. Deepfakes, explained. MIT Sloan. 21.7.2020. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained. 20.11.2020.

Tietosuojalaki 1050/2018.

TMBDF. 2022. [GUIDE] - DeepFaceLab 2.0 Guide. https://mrdeepfakes.com/forums/thread-guide-deepfacelab-2-0-guide. 8.2.2022.

Torzdf. 2022a. [Guide] Extraction - A Workflow. Faceswap Forum. https://forum.faceswap.dev/viewtopic.php?t=27. 24.9.2022.

Torzdf. 2022b. [Guide] Training in Faceswap. Faceswap Forum. https://forum.faceswap.dev/viewtopic.php?t=146. 24.9.2022.

Turek, M. 2021. Media Forensics (MediFor). Defense Advanced Research Projects Agency. https://www.darpa.mil/program/media-forensics. 28.4.2021.

Twitter. 2022. Synthetic and manipulated media policy. https://help.twitter.com/en/rules-and-policies/manipulated-media. 24.2.2022.

Uchill, J. 2019. Why the deepfakes threat is shallow. Axios. 15.8.2019. https://www.axios.com/why-the-deepfakes-threat-is-shallow-16caf6a0-af83-4dbc-9008-6a2d4a2f08ae.html. 15.2.2021

Uudet lukutaidot. 2022a. Medialukutaito. https://uudetlukutaidot.fi/medialukutaito/. 21.2.2022.

Uudet lukutaidot. 2022b. Median tulkinta ja arviointi. https://uudetlukutaidot.fi/medialukutaito/median-tulkinta-ja-arviointi/. 21.2.2022.

Vehkoo, J. 2021. Valheenpaljastaja: Näin tunnistat videohuijauksen. Yle. 17.1.2021. https://yle.fi/aihe/artikkeli/2021/01/17/valheenpaljastaja-nain-tunnistat-videohuijauksen. 28.2.2022.

Mediakasvatusvideo

Opinnäytetyön toiminnallisen osan tuloksena syntynyt mediakasvatusvideo: https://youtu.be/ByTdr45K_pQ.

Vaihtoehtoinen linkki mediakasvatusvideoon: https://youtu.be/sJLwIXU6u5E.