

Edith Cowan University  
**Research Online**

---

Theses : Honours

Theses

---

2019

## Facial re-enactment, speech synthesis and the rise of the Deepfake

Nicholas Gardiner  
*Edith Cowan University*

Follow this and additional works at: [https://ro.ecu.edu.au/theses\\_hons](https://ro.ecu.edu.au/theses_hons)



Part of the [Artificial Intelligence and Robotics Commons](#), and the [Music Commons](#)

---

### Recommended Citation

Gardiner, N. (2019). *Facial re-enactment, speech synthesis and the rise of the Deepfake*.  
[https://ro.ecu.edu.au/theses\\_hons/1530](https://ro.ecu.edu.au/theses_hons/1530)

This Thesis is posted at Research Online.  
[https://ro.ecu.edu.au/theses\\_hons/1530](https://ro.ecu.edu.au/theses_hons/1530)

# Edith Cowan University

## Copyright Warning

You may print or download ONE copy of this document for the purpose of your own research or study.

The University does not authorize you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site.

You are reminded of the following:

- Copyright owners are entitled to take legal action against persons who infringe their copyright.
- A reproduction of material that is protected by copyright may be a copyright infringement. Where the reproduction of such material is done without attribution of authorship, with false attribution of authorship or the authorship is treated in a derogatory manner, this may be a breach of the author's moral rights contained in Part IX of the Copyright Act 1968 (Cth).
- Courts have the power to impose a wide range of civil and criminal sanctions for infringement of copyright, infringement of moral rights and other offences under the Copyright Act 1968 (Cth). Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

---

# FACIAL RE-ENACTMENT, SPEECH SYNTHESIS AND THE RISE OF THE DEEPPFAKE

By Nicholas Gardiner

This thesis is presented in partial fulfilment of the degree of

**Bachelor of Music Honours**

Western Australian Academy of Performing Arts

Edith Cowan University

2019

---

---

## Copyright declaration

I certify that this thesis does not, to the best of my knowledge and belief:

1. incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher degree or diploma in any institution of higher education;
2. contain any material previously published or written by another person except where due reference is made in the text of this thesis;
3. contain any defamatory material;
4. contain any data that has not been collected in a manner consistent with ethics approval.

Signed:



Date: **8/11/18**

This copy is the property of Edith Cowan University. However, the literary rights of the author must also be respected. If any passage from this thesis is quoted or closely paraphrased in a paper or written work prepared by the user, the source of the passage must be acknowledged in the work. If the user desires to publish a paper or written work containing passages copied or closely paraphrased from this thesis, which passages would in total constitute an infringing copy for the purpose of the Copyright Act, he or she must first obtain the written permission of the author to do so.

---

## ABSTRACT

Emergent technologies in the fields of audio speech synthesis and video facial manipulation have the potential to drastically impact our societal patterns of multimedia consumption. At a time when social media and internet culture is plagued by misinformation, propaganda and “fake news”, their latent misuse represents a possible looming threat to fragile systems of information sharing and social democratic discourse. It has thus become increasingly recognised in both academic and mainstream journalism that the ramifications of these tools must be examined to determine what they are and how their widespread availability can be managed.

This research project seeks to examine four emerging software programs – *Face2Face*, *FakeApp*, *Adobe VoCo* and *Lyrebird* – that are designed to facilitate the synthesis of speech and manipulate facial features in videos. I will explore their positive industry applications and the potentially negative consequences of their release into the public domain. Consideration will be directed to how such consequences and risks can be ameliorated through detection, regulation and education. A final analysis of these three competing threads will then attempt to address whether the practical and commercial applications of these technologies are outweighed by the inherent unethical or illegal uses they engender, and if so; what we can do in response.

## TABLE OF CONTENTS

Glossary	3
1. Introduction	7
1.1. The rise of the deepfake	7
1.2. 21 <sup>st</sup> century misinformation	8
1.3. The dilemma of digital avatars	10
2. Technology & applications	13
2.1. Facial re-rendering	13
2.1.1. <i>Face2Face</i>	13
2.1.2. <i>FakeApp</i>	15
2.2. Speech synthesis	15
2.2.1. <i>Adobe VoCo</i>	15
2.2.2. <i>Lyrebird</i>	16
2.3. Other research	17
2.4. Applications	18
2.4.1. Multimedia industries	19
2.4.2. Broader social applications	20
2.4.3. Creative uses	21
3. Consequences	22
3.1. Social democratic discourse and information sharing	22
3.2. Personal image and avatar appropriation	26
3.3. Fraud and legal recourse	28
4. Mitigation	30
4.1. Detection	31
4.1.1. Multimedia forensics	31
4.1.2. Watermarking	33

4.1.3. Neural networks	35
4.2. Regulation	37
4.2.1. Government	37
4.2.2. Legal avenues	39
4.2.3. Other authorities	41
4.3. Education	42
5. Analysis	45
5.1. A fake news apocalypse?	45
5.2. The inevitable spread of open-source software	46
5.3. A crisis of confidence	48
5.4. A multi-sector response	50
6. Conclusions	52
7. Reference List	55
APPENDIX A	
Interviews	69
Developer - Justus Thies	69
Industry Professional - Joshua Hogan	71
APPENDIX B	
Consent Forms	

## GLOSSARY

**Artificial intelligence (AI)** – a form of machine intelligence where the machine mimics cognitive functions usually associated with human minds, such as recognition and learning

**Astroturfing** – the practice of masking the sponsors of a message or organization (e.g. political, advertising, religious or public relations) to make it appear as though it originates from and is supported by grassroots participants

**Automatic dialog replacement (ADR)** – the process of re-recording the dialogue of the original actor after the filming process to improve audio quality or reflect dialogue/scene changes

**Convolutional neural network (CNN)** – a class of deep learning artificial neural network focussed on image recognition and analysis

**Confirmation bias** – the tendency to search for, interpret, favour and recall information in a way that confirms one's pre-existing beliefs or hypotheses

**Computer generated imagery (CGI)** – application of computer graphics to create or contribute to images, such as film, video, video games etc.

**Deep learning** – a machine learning method based on learning data representations, such as speech, audio and image recognition

**Deepfake** – a portmanteau of “deep learning” and “fake”, commonly associated with an AI generated machine learning technique of digitally swapping facial architecture

**Digital avatar** – a graphical representation of a character or person. In the context of this dissertation, the representation is life-like in both speech and visual appearance

**Digital rights management (DRM)** – a set of access control technologies for restricting the use of proprietary hardware and copyrighted works

**Facial re-enactment** – a method of facial performance capture whereby the facial expressions of a source “actor” are transferred and re-rendered to the facial architecture of a video target “actor” in a photo-realistic fashion

**Facial replacement** – a method of replacing target and source actor facial features that is commonly associated with a “deepfake”



**Filter bubble** – the intellectual isolation that can occur when websites make use of algorithms to selectively assume the information a user would want to see to the exclusion of contradictory viewpoints or facts

**Fourth estate** – a reference to the press and news media, both in their explicit capacity of advocacy and reporting on social causes and their implicit ability to frame political issues

**Generative adversarial network (GAN)** – a class of AI algorithms used in unsupervised machine learning, implemented by having a system of two artificial neural networks contest each other in a zero-sum game

**Image forensics** – the academic analysis of image files to evaluate the presence of forgeries or digital manipulation thus determining the authenticity of the image and its content

**Machine learning** – a field of computer science that uses statistical techniques to give computer systems the ability to “learn” with data, as opposed to task-specific algorithms or explicit programming

**Multimedia forensics** – the study of ensuring authenticity, origin and provenance of an image or video without the help of an embedded security scheme

**Open source software** – source code for software that anyone can inspect, modify and enhance, allowing open collaboration and editing by the general public

**Right to oblivion** – the right of individuals to determine the development of their life in an autonomous way, without being perpetually or periodically stigmatized as a consequence of a specific action performed in the past, commonly associated with electronically archived internet history

**Source actor** – the human subject whose features or facial architecture are projected onto the target actor

**Speech enhanced generative adversarial network (SEGAN)** – a form of generative adversarial network dedicated to end-to-end speech enhancement

**Speech synthesis** – the artificial production of human speech via hardware or software systems and programs

**Target actor** – the human subject that is the target of a digital facial manipulation, via either facial re-enactment or facial replacement

**Text-to-speech** – a software or hardware system converts standard language text into synthesised speech

**Watermarking** – the insertion of a hidden digital signature into a software, signal or output data, allowing others to verify its authenticity or integrity, or determine its owner

## CHAPTER 1 – INTRODUCTION

*"If everything is real .... then nothing is real as well"*

*- David Lynch*

### ***1.1 The rise of the deepfake***

In late 2017, an anonymous user under the pseudonym of "Deepfakes" posted several pornographic videos purporting to feature famous actresses such as Gal Gadot, Scarlett Johansson, Maisie Williams, Taylor Swift and Aubrey Plaza on the popular media aggregation site Reddit<sup>1</sup>. Though the videos were quickly debunked, the peculiar deep learning method of facial replacement quickly spread throughout various subreddits and internet forums, gaining widespread notoriety and media attention.<sup>2</sup> On February 7<sup>th</sup> 2018, all subreddits' relating to this popular technique known as "deepfaking"<sup>3</sup> were banned,<sup>4</sup> with further removal enacted on other multimedia sites, such as *Twitter*, *Pornhub*, *Gfycat* and *Discord*.<sup>5</sup> Despite this, deepfaked content has continued to propagate throughout the worldwide web today.<sup>6</sup>

---

<sup>1</sup> [www.reddit.com](http://www.reddit.com) is a social news aggregation, web content rating and discussion website where users submit content in the form of links, text or images. Posts are organised by subject into 'subreddits' created by users and cover a variety of topics within the terms of use.

<sup>2</sup> Samantha Cole, "A.I-Assisted Fake Porn Is Here and We're All Fucked," Motherboard, last modified December 11, 2017, [https://motherboard.vice.com/en\\_us/article/gdydym/gal-gadot-fake-ai-porn](https://motherboard.vice.com/en_us/article/gdydym/gal-gadot-fake-ai-porn).

<sup>3</sup> An internet colloquialism, named after the user who developed the process and defined in the glossary

<sup>4</sup> <https://www.reddit.com/r/deepfakes/>; the ban coincided with an update to Reddit's terms of use disallowing any subreddits that featured "involuntary pornography" including depictions that were faked (<https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-post-involuntary-pornography>). The Reddit admin team also banned several other subreddits which did not feature pornography, but which were designed to support the deepfake process (see <https://www.reddit.com/r/fakeapp>) and any older forums not related to deepfaking but which breached the new terms of use. It can be inferred that the terms of use update was due to the widespread emergence and reporting on deepfaking.

<sup>5</sup> Janko Roettgers, "Porn Producers Offer to Help Hollywood Take Down Deepfake Videos," Variety, last modified February 21, 2018, <https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/>.

<sup>6</sup> Samantha Cole, "Reddit Just Shut Down the Deepfakes Subreddit," Motherboard, last modified February 7, 2018, [https://motherboard.vice.com/en\\_us/article/neqb98/reddit-shuts-down-deepfakes](https://motherboard.vice.com/en_us/article/neqb98/reddit-shuts-down-deepfakes).

Concerning the above events, it is important to note that the initial developer of deepfaking was a software engineer.<sup>7</sup> However, the tools and functions utilised in its creation<sup>8</sup> were open sourced from software companies such as *Google* and *Nvidia*. In other words, the development process required a technical know-how of computational parameters and algorithms, but the programs employed were publicly available for widespread use. Subsequent developments of self-contained applications and tutorials,<sup>9</sup> such as *FakeApp*<sup>10</sup>, have simplified and automated many aspects of the creation process, making the technique simpler and thus more accessible to a wider audience.<sup>11</sup> The inception of deepfakes as a tool to primarily produce pornographic content and the open source nature in which it emerged play an important contextual role to the questions raised in this thesis.

### ***1.2 21<sup>st</sup> century misinformation***

Throughout the early 21<sup>st</sup> century, our society as content consumers has evolved a certain dependence on the internet and digital information distribution to supplement our formed opinions, biases and outlook. In many respects, social media sites like *Facebook* and *Twitter*, and multimedia services like *YouTube*, have surpassed conventional news media outlets,<sup>12</sup> allowing near instantaneous propagation of video and audio content across

---

<sup>7</sup> Cole, "A.I.-Assisted Fake Porn".

<sup>8</sup> A series of python scripts in a deep learning algorithm. The application is based on a class of neural network know as a generative adversarial network (GAN). For more information, see A.I. Wiki, "A Beginner's Guide to Generative Adversarial Networks (GANs)," Skymind, last accessed November 6, 2018, <https://skymind.ai/wiki/generative-adversarial-network-gan>.

<sup>9</sup> deepfakesclub, "DeepFakes FakeApp Tutorial," DeepFakes.Club, last accessed April, 2018, <https://www.deepfakes.club/tutorial/>. This website is no longer available.

<sup>10</sup> FakeApp Ver. 2.2, deepfakeapp, <https://www.fakeapp.org/>. This website is no longer available

<sup>11</sup> Alex Hern, "My May-Thatcher deepfake won't fool you but its tech may change the world," The Guardian, last modified March 12, 2018, <https://www.theguardian.com/technology/2018/mar/12/may-thatcher-deepfake-face-swap-tech-change-world>.

<sup>12</sup> Monica Anderson and Andrea Caumont, "How social media is reshaping news," Pew Research, last modified September 24, 2014, <http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/>. For further information, see Jane Wakefield, "Social media 'outstrips T.V.' as news source for young people," British Broadcasting Corporation (BBC), last modified June 15, 2016, <http://www.bbc.com/news/uk-36528256>.

multiple online platforms. Search engines such as *Google* have further amplified the rise of the internet as our primary informational network.

However, more pressingly, we live in a time when social media and internet culture is plagued by misinformation, propaganda and “fake news” spread throughout the worldwide web by state actors, third parties and end users. Our systems of information-sharing and social democratic discourse remain fragile and weathered by recent world events such as the Cambridge Analytica scandal,<sup>13</sup> Russian interference in the 2016 US elections<sup>14</sup> and the unfolding fake news phenomena.<sup>15</sup>

Emergent technologies and software applications in the fields of audio speech synthesis and video facial manipulation, such as *FakeApp*, have the potential to drastically exacerbate these ongoing issues, warping and distorting our societal patterns of multimedia consumption in detrimental and unmeasured ways. The widespread public availability of such products used to digitally alter multimedia content, so as to misrepresent or misportray a target actor in a negative way, bodes ominously for our cultural harmony and systems of governance. Thus, it is imperative that these technologies be examined to determine the ethical and legal implications of their introduction into the public sphere.

---

<sup>13</sup> ABC, “Cambridge Analytica harvested data from more than 87 million Facebook users, whistleblower says,” Australian Broadcasting Corporation (ABC), last modified April 18, 2018, <http://www.abc.net.au/news/2018-04-18/cambridge-analytica-employee-testifies-before-uk-committee/9670192>. For further information, see Craig Timberg and Karla Adam, “I’m not going to be bullied by Facebook’: Cambridge Analytica whistleblower tells his story,” Sydney Morning Herald, last modified March 23, 2018, <https://www.smh.com.au/technology/i-m-not-going-to-be-bullied-by-facebook-cambridge-analytica-whistleblower-tells-his-story-20180322-p4z5sa.html>.

<sup>14</sup> Director of National Intelligence (D.N.I.), “Background to “Assessing Russian Activities and Intentions in Recent US Elections”: The Analytic Process and Cyber Incident Attribution,” ([https://www.dni.gov/files/documents/ICA\\_2017\\_01.pdf](https://www.dni.gov/files/documents/ICA_2017_01.pdf): Office Of The Director Of National Intelligence, 2017).

<sup>15</sup> Elle Hunt, “What is fake news? How to spot it and what you can do to stop it,” The Guardian, last modified December 18, 2016, <https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>.

### *1.3 The dilemma of digital avatars*

This dissertation will address the following questions:

1. What are the ethical and legal implications of audio speech synthesis and video facial manipulation?
2. Are the positive applications of these technologies outweighed by the unethical and illegal uses they engender?
3. If so, what actions can be taken in response to these consequences?

I will primarily examine the following emerging software applications and research projects:

1. *Face2Face*<sup>16</sup> (facial re-enactment)
2. *FakeApp*<sup>17</sup> (facial replacement)
3. *Adobe VoCo*<sup>18</sup> (speech synthesis)
4. *Lyrebird*<sup>19</sup> (speech synthesis)

In essence, the interaction of facial manipulation with accompanying speech raises the spectre of creating re-visualised **digital avatars** of target actors, allowing any lay person to manipulate video and audio content for their own purposes. Viewed in the context of recent controversies surrounding deepfaking and the issues of digital misinformation discussed in Sections 1.1 and 1.2, this will be the main focus of this dissertation.

Chapter 2 will focus on the research and development underlying these technologies while citing public exhibitions of their potential positive applications. Commercial and

---

<sup>16</sup> Matthias Niessner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)," (<https://www.youtube.com/watch?v=ohmajJTcpNk>: YouTube, 2016).

<sup>17</sup> FakeApp.

<sup>18</sup> Adam Finkelstein, "VoCo: Text-based Insertion and Replacement in Audio Narration," (<https://www.youtube.com/watch?v=RB7upq8nzIU>: YouTube, 2017).

<sup>19</sup> Lyrebird Ver. Beta, Lyrebird, <https://lyrebird.ai/>.

practical uses will be discussed and enlightened with supplemental commentary from interviews with industry figures and developers. I will focus on the film, game and advertising industries, which are the primary beneficiaries of these tools, but also briefly explore the creative arts and broader social applications.

Chapter 3 will hypothesise the unethical and illegal implications inherent in their potential to create digital avatars. These will be classified under three areas:

1. Social democratic discourse and information sharing
2. Personal image and avatar appropriation
3. Fraud and legal recourse

Chapter 4 will analytically review various ways to limit the negative aspects inherent in the design and potential of these products. Specifically, I will focus on three key areas of mitigation my research has uncovered:

1. Detection via multimedia forensics, watermarking and neural networks
2. Regulation through Government/legal intervention and other authorities
3. Education of the public in the form of information campaigns and gradual introduction of the software

Each of these will be critically evaluated to hypothesise their effectiveness in light of various sociocultural studies.

The final chapter will draw the competing threads of Chapters 2, 3 and 4 together to qualitatively weigh the various issues and concerns apparent in their widespread availability. In the context of the deepfake phenomena, historical reflection and current open source developments, I will evaluate whether such technological advancements can be prevented from manifesting in the public sphere, and if not; on what terms that emergence would best occur.

The issues canvassed in this paper are highly emergent, with fields of study and technology development moving exceedingly quickly. I have attempted to provide numerous literature and media references, with additional industry interviews, to track and highlight the substantive developments in the areas discussed. Nevertheless, this dissertation aims to capture the discussion at this point in time, however rapidly that may change.



## CHAPTER 2 – TECHNOLOGY & APPLICATIONS

This chapter will focus on background research surrounding these technologies. It is intended to inform the reader of their basic functions and intended applications in order to understand their potential.

Exploration of these emergent digital manipulation technologies will initially be separated into two distinct topics; facial re-rendering (comprising facial re-enactment, replacement and other derivative uses) and speech synthesis. Whilst these are separate areas of research and development, with differing fields of application, their potential when combined allows anyone to fabricate manipulated video with accompanying speech, thereby creating their own digital avatars. Therefore, I will also briefly canvass other academic sources and research which reinforce the interaction between the video and audio aspects of the above technologies through various iterations of lip syncing and dialog/speech alignment.

Finally, I will explore their intended applications, primarily focussing on the multimedia content creation industries, but also noting their creative uses and positive benefits in other areas of society. Hopefully, this will elucidate the idea these technologies are not without great benefit to the advancement of human cultural and scientific progress, even in limited cases like deepfaking.

### ***2.1 Facial re-rendering***

#### 2.1.1 Face2Face (facial re-enactment)

As defined in the glossary, facial re-enactment represents a form of gesture manipulation, adjusting the target actor's eyes, mouth, nose, forehead and jaw in a video output to reflect those of the source actor.<sup>20</sup> Although research into different forms of facial re-enactment

---

<sup>20</sup> Niessner, "Face2Face."

has been on-going for years,<sup>21,22</sup> the introduction of *Face2Face* by Justus Thies<sup>23</sup> in 2016 and expanded on in his PhD<sup>24</sup> represented a large step forward in commoditising<sup>25</sup> the product for consumer use. Its demonstration of online real-time re-enactment<sup>26</sup> using a standard webcam allowed for convincing and instantaneous re-rendering of the target actor with a relatively simple home setup. This was further expanded to include gesture driven facial re-enactment using hands and an inertial measurement unit<sup>27</sup>, and more recently *HeadOn*<sup>28,29</sup> and *Deep Video Portraits*<sup>30,31</sup>; portrait driven transfers of torso, head motion, face expression, and eye gaze. It is thus easy to see how quickly developments have been made in this area. And while it extends beyond the limits of this paper, we will likely see full body re-enactment<sup>32</sup> videos<sup>33</sup> in the near future.

---

<sup>21</sup> Michael Zollhöfer et al., "Real-time non-rigid reconstruction using an RGB-D camera," *ACM Transactions on Graphics (TOG)* 33, no. 4 (2014).

<sup>22</sup> Justus Thies et al., "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics (TOG)* 34, no. 6 (2015).

<sup>23</sup> Justus Thies et al., "Face2Face: Real-time face capture and reenactment of RGB videos," In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, Nevada: Institute of Electrical and Electronics Engineers (IEEE), 2016),

<sup>24</sup> Justus Thies, "Face2Face: Real-time Facial Reenactment," (PhD dissertation, Friedrich Alexander University, 2017).

<sup>25</sup> Christian Theobalt et al., "Real-time Expression Transfer for Facial Reenactment," (US Patent App. 15/256,710, 2018).

<sup>26</sup> Online real-time refers to live gesture animations with a source actor present, as opposed to pre-generated and rendered video output. In other words, live re-enacting rather than a video prepared earlier.

<sup>27</sup> Justus Thies, Michael Zollhöfer, and Matthias Nießner, "IMU2Face: Real-time Gesture-driven Facial Reenactment," *arXiv preprint arXiv:1801.01446* (2017).

<sup>28</sup> Justus Thies et al., "HeadOn: Real-time Reenactment of Human Portrait Videos," *ACM Transactions on Graphics (TOG)* 37, no. 4 (2018).

<sup>29</sup> Christian Theobalt, "HeadOn: Real-time Reenactment of Human Portrait Videos - SIGGRAPH 2018," (<https://www.youtube.com/watch?v=KRllyxqsBTM>: YouTube, 2018).

<sup>30</sup> Hyeonwoo Kim et al., "Deep Video Portraits," *ACM Transactions on Graphics (TOG)* 37, no. 4 (2018). For further information, see George Dvorsky, "Deepfake Videos Are Getting Impossibly Good," Gizmodo, last modified June 14, 2018, <https://www.gizmodo.com.au/2018/06/deepfake-videos-are-getting-impossibly-good/>.

<sup>31</sup> Christian Theobalt, "Deep Video Portraits - SIGGRAPH 2018," (<https://www.youtube.com/watch?v=qc5P2bvfl44>: YouTube, 2018).

<sup>32</sup> Caroline Chan et al., "Everybody Dance Now," *arXiv Preprint arXiv:1808.07371* (2018). For further information, see Peter Farquhar, "An AI program will soon be here to help your deepfake dancing - just don't call it deepfake," Business Insider Australia, last modified August 27, 2018, <https://www.businessinsider.com.au/artificial-intelligence-ai-deepfake-dancing-2018-8>.

<sup>33</sup> Caroline Chan, "Everybody Dance Now," (<https://www.youtube.com/watch?v=PCBTZh41Ris>: YouTube, 2018). See also Two Minute Papers, "Everybody Dance Now," (<https://www.youtube.com/watch?v=cEBgi6QYDhQ>: YouTube, 2018).

### 2.1.2 FakeApp (facial replacement / deepfake)

*FakeApp* and other similar applications<sup>34</sup> are publicly developed open-source software with no associated academic literature or private backing. Their method of facial replacement uses a generative adversarial network (GAN)<sup>35</sup> to “train” AI models against corresponding face datasets. The process takes several hours and decent computer know-how but can be accomplished on a consumer brand home computer with a decent graphics card and the appropriate pre-requisite software, such as *FakeApp*. Although the applications have been used extensively to create fake celebrity pornography, they have also been utilised in less ethically unconscionable scenarios such as replacing faces in movies with the famous actor Nicholas Cage<sup>36</sup> or replacing human CGI with machine learnt deepfakes.<sup>37</sup> As will become apparent later in this paper, the nature of *FakeApp*'s use in the public sphere is important to compare with the other unreleased technologies.

## ***2.2 Speech synthesis***

### 2.2.1 Adobe VoCo (speech synthesis)

Although classified as speech synthesis, *VoCo* is more accurately a two-step process of text-to-speech and voice conversion<sup>38</sup>. It features text-based editing, alternative synthesis of words and manual editing of length, stitching points, pitch profile and amplitude,<sup>39</sup>

---

<sup>34</sup> OpenFaceSwap Ver. 0.9, Deepfakesclub, <https://www.deepfakes.club/openfaceswap-deepfakes-software/>. This website is no longer active but linked to an alternative face swap application that utilised similar methods to *FakeApp*.

<sup>35</sup> Ian Goodfellow et al., "Generative adversarial nets," In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, (Cambridge, Massachusetts: MIT Press, 2014),

<sup>36</sup> Nick Cage Deepfakes, "Nic Cage deepfakes mini compilation," (<https://www.youtube.com/watch?v=2jp4M1clJ5A>: YouTube, 2018).

<sup>37</sup> derpfakes, "Grand Moff Tarkin | Derpfakes," (<https://www.youtube.com/watch?v=6Zn1vt9vdwU&feature=youtu.be>: YouTube, 2018).

<sup>38</sup> Zeyu Jin et al., "CUTE: A concatenative method for voice conversion using exemplar-based unit selection," In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Shanghai, China: Institute of Electrical and Electronics Engineers (IEEE), 2016), This paper presents the original voice conversion techniques which were subsequently used to create *VoCo*.

<sup>39</sup> Finkelstein, "VoCo: Text-based Insertion and Replacement in Audio Narration."

thereby allowing users to tweak and refine generated words in a more accurate portrayal of human voice patterns. Originally presented at *Adobe MAX 2016 (Sneak Peeks)*<sup>40</sup> as a collaboration between Princeton University students and Adobe Research,<sup>41</sup> *VoCo* is not publicly available and in all likelihood has been shelved by Adobe due to ethical and legal concerns<sup>42</sup> not dissimilar from those raised in this thesis. Irrespective of this, it provides important contextual comparison to the development of *Lyrebird*.

### 2.2.2 Lyrebird (speech synthesis)

*Lyrebird* is the name of Canadian start-up company with their title product speech synthesiser in beta.<sup>43</sup> Researched initially at the University of Montreal under the moniker *Char2Wav*,<sup>44</sup> the developers are notable for having produced a synthesised photo-realistic lip-sync of former President Obama with corresponding synchronized speech.<sup>45,46</sup> In mid-2018, *Lyrebird* began offering, via their website, approved applicants a personalised speech synthesiser of a consenting recorded participant,<sup>47</sup> marking a small step towards widespread availability. Per their ethical statement of intent,<sup>48</sup> this confirms their plans to

---

<sup>40</sup> Adobe Creative Cloud, "#VoCo. Adobe MAX 2016 (Sneak Peeks) | Adobe Creative Cloud," (<https://www.youtube.com/watch?v=I3l4XLZ59iw>: YouTube, 2016).

<sup>41</sup> Zeyu Jin et al., "VoCo: text-based insertion and replacement in audio narration," *ACM Transactions on Graphics (TOG)* 36, no. 4 (2017).

<sup>42</sup> dhermanq, "Beta Testing #VoCo," Adobe, last modified November 8, 2016, <https://forums.adobe.com/thread/2233703>.

<sup>43</sup> James Vincent, "Lyrebird claims it can recreate any voice using just one minute of sample audio," *The Verge*, last modified April 24, 2017, <https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird>.

<sup>44</sup> Jose Sotelo et al., "Char2wav: End-to-end speech synthesis," (Paper presented at the International Conference on Learning Representations, Toulon, France, February 17, 2017).

<sup>45</sup> Rithesh Kumar et al., "ObamaNet: Photo-realistic lip-sync from text," (Paper presented at the 31st Conference on Neural Information Processing Systems, Long Beach, California, December 6, 2017). See also Rithesh Kumar, "Lyrebird - Create a digital copy of your voice," Montreal Institute for Learning Algorithms & Lyrebird, last modified September 4, 2018, <http://ritheshkumar.com/obamanet/>.

<sup>46</sup> Lyrebird, "Lyrebird - Create a digital copy of your voice.," ([https://www.youtube.com/watch?v=YfU\\_sWHT8mo](https://www.youtube.com/watch?v=YfU_sWHT8mo): YouTube, 2017).

<sup>47</sup> Lyrebird, "Vocal Avatar API," Lyrebird, last modified 2018, <https://lyrebird.ai/vocal-avatar-api>.

<sup>48</sup> Lyrebird, "With great innovation comes great responsibility", *Ethics (blog)*, Lyrebird, 2017, <https://lyrebird.ai/ethics/>.

introduce their speech synthesiser incrementally, allowing the public time to understand and adapt to the changing digital environment.

### ***2.3 Other research***

The fields of computer science and AI deep learning development are continually moving forward in new and unexpected ways. The above-mentioned technologies represent only a fraction of the work being expounded on in various fields of study around the world. My research into this area has uncovered a number of different papers which reinforce the link between re-visualised avatar production and audio dialog generation, allowing the generation of full-fledged digital avatars. These include:

- facial landmark tracking<sup>49</sup>
- facial expression mapping<sup>50</sup>
- voice cloning<sup>51,52</sup>
- audio speech generation<sup>53,54</sup>
- audio driven facial re-enactment<sup>55</sup>
- reconstructed lip-syncing<sup>56</sup>

---

<sup>49</sup> Li Wing Yee and Dirk Scheiders, "Facial Landmark Tracking Final Report," (Semantics Scholar: Allen Institute for Artificial Intelligence, 2017).

<sup>50</sup> Katsuhiro Suzuki et al., "Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display," (Paper presented at the 2017 IEEE Virtual Reality (VR), Los Angeles, California, March 18-22, 2017).

<sup>51</sup> Sercan O Arik et al., "Neural Voice Cloning with a Few Samples," *arXiv preprint arXiv:1802.06006* (2018). For further information, see Edd Gent, "A.I. hears snippets of you, then clones your voice," *New Scientist* 237, no. 3167 (2018).

<sup>52</sup> Yang Gao, Rita Singh, and Bhiksha Raj, "Voice Impersonation using Generative Adversarial Networks," *arXiv preprint arXiv:1802.06840* (2018).

<sup>53</sup> Aaron van den Oord et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499* (2016).

<sup>54</sup> Chris Donahue, Julian McAuley, and Miller Puckette, "Synthesizing Audio with Generative Adversarial Networks," *arXiv preprint arXiv:1802.04208* (2018).

<sup>55</sup> Seyed Ali Jalalifar, Hosein Hasani, and Hamid Aghajan, "Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks," *arXiv preprint arXiv:1803.07461* (2018).

<sup>56</sup> Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (TOG)* 36, no. 4 (2017). See also David Bunker,

- speech enhancement<sup>57</sup>
- language transfer<sup>58</sup>
- static image animation<sup>59,60</sup>; and
- 3D face modelling<sup>61,62</sup>

Thus, it should be clear from the above that given a small sample of video and speech, the ability to create photorealistic and controllable digital avatars of anyone, with accompanying lip-synced high-quality dialog, currently exists in the world today. Such potential would lead one to ask how this can benefit the wider community.

## ***2.4 Applications***

Although mostly unavailable to the public at this time, the above-mentioned technologies have the potential to revolutionise various aspects of multimedia content creation. I will briefly focus on three areas:

- Multimedia industries (particularly film, game and advertising)
- Broader social applications

---

"Speech2Face: Reconstructed Lip Syncing with Generative Adversarial Networks," In *Data Reflexions: Thoughts and Projects* (2017), [https://www.dbunker.io/docs/2017\\_Bunker\\_Speech2FaceProposal.pdf](https://www.dbunker.io/docs/2017_Bunker_Speech2FaceProposal.pdf).

<sup>57</sup> Jaime Lorenzo-Trueba et al., "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," *arXiv preprint arXiv:1803.00860* (2018).

<sup>58</sup> Santiago Pascual et al., "Language and Noise Transfer in Speech Enhancement Generative Adversarial Network," *arXiv preprint arXiv:1712.06340* (2017).

<sup>59</sup> Hadar Averbuch-Elor et al., "Bringing portraits to life," *ACM Transactions on Graphics (TOG)* 36, no. 6 (2017). See also Hadar Elor, "Bringing Portraits to Life " (<https://www.youtube.com/watch?v=RetOjL1Fhw>: YouTube, 2018).

<sup>60</sup> Hai X Pham, Yuting Wang, and Vladimir Pavlovic, "Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network," *arXiv preprint arXiv:1803.07716* (2018).

<sup>61</sup> Aaron S Jackson et al., "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression," In *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy: Institute of Electrical and Electronics Engineers (IEEE), 2017), For further information, see James Vincent, "Make a 3D model of your face from a single photo with this A.I. tool," The Verge, last modified September 18, 2017, <https://www.theverge.com/2017/9/18/16327906/3d-model-face-photograph-ai-machine-learning>.

<sup>62</sup> Yao Feng et al., "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network," *arXiv preprint arXiv:1803.07835* (2018).

- Creative uses

#### 2.4.1 Multimedia industries

As logic dictates, industries that utilise digital video characters stand to gain the most from re-synthesising video and audio, primarily by lowering barriers to access for high quality visual effects. This is particularly true of the film<sup>63</sup>, game<sup>64</sup> and advertising<sup>65</sup> industries, which express hopeful enthusiasm at their potential availability. As Josh Hogan noted, the ease of editing and synthetically replacing actor's dialog or other aspects of their performance would mean cost reductions and time saved in the editing and post-production spheres,<sup>66</sup> but may also lead to other pitfalls – which I discuss in Section 3.2.

Pablo Garrido<sup>67</sup> and Thies<sup>68</sup> noted in the early iterations of facial re-enactment that it has huge potential for redubbing films and advertising into different languages, allowing perfectly aligned foreign dialog with corresponding mouth movements. Further benefits are also noted in areas of video game animation and Virtual/Augmented Reality<sup>69</sup> applications. As Zeyu Jin<sup>70</sup> explains, positive applications of voice synthesis are primarily in the fields of ADR and voice narration of video games, demonstration videos, documentaries and podcasts.

Finally, in spite of recent history, even face replacement applications like *FakeApp* can serve a purpose in the entertainment industry by, for example, transferring facial

---

<sup>63</sup> Jon Fusco, "Is Adobe's Project VoCo the Photoshop for Audio?," In *No Film School* (2016), <https://nofilmschool.com/2016/11/adobe-project-voco>.

<sup>64</sup> Lisa Torekull et al., "The Future Of Storytelling And Game Writing," In *Future of Media* (2017), <http://fom.csc.kth.se/archive/files/2016-Gaming/08-Omnius-Chapter.pdf>.

<sup>65</sup> Justus Thies, Interviewed by author via email correspondence, Perth, September 25, 2018.

<sup>66</sup> Josh Hogan, Interviewed by author via email correspondence, Perth, September 25, 2018.

<sup>67</sup> Pablo Garrido et al., "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track," *Computer Graphics Forum* 34, no. 2 (2015).f

<sup>68</sup> Thies, "Face2Face."

<sup>69</sup> Justus Thies et al., "FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality," *arXiv preprint arXiv:1610.03151* (2016).

<sup>70</sup> Jin et al., "VoCo."

architecture onto deceased or younger film characters,<sup>71</sup> or simply replacing faces for creative purposes. Similarly, Christopher Buccafusco further notes that digital reanimation may aid in film restoration, helping to preserve cultural artistic heritage and historical works.<sup>72</sup>

#### 2.4.2 Broader social applications

As noted by Thomas Giger, speech synthesis has huge potential for the radio industry, streamlining news bulletin systems and traffic reports with automated voiceovers and low-cost vocalisations.<sup>73</sup> Wesley Mattheyses details usage for both in areas such as personalised digital or virtual assistants, remote teaching applications and speech therapy,<sup>74</sup> particularly video conferencing<sup>75</sup> and live language translation<sup>76</sup>.

In terms of medical technology, Thies noted the potential benefits of their research for psychology<sup>77</sup>, surgery<sup>78</sup> and other educational training applications, allowing trainees to analyse and work from digital representations of people. Comparably, the *Lyrebird* team are actively demonstrating positive social change, partnering with The ALS Association<sup>79</sup> to develop Project Revoice.<sup>80</sup> This joint venture provides the opportunity for sufferers of ALS<sup>81</sup> to reclaim their own unique voice, albeit in computerised form, even after losing the

---

<sup>71</sup> derpfakes, "Solo | A Derpfakes Story," (<https://www.youtube.com/watch?v=ANXucrz7Hjs&feature=youtu.be>: YouTube, 2018).

<sup>72</sup> Christopher Buccafusco, Jared Vasconcellos Grubow, and Ian Postman, "Preserving Film Preservation from the Right of Publicity," In *2018 Cardozo Law Review de novo 1 Cardozo Legal Studies Research Paper No. 544* (2018), <https://ssrn.com/abstract=3145533>

<sup>73</sup> Thomas Giger, "#VoCo For Radio: Revolutionary Tool Or Complete #NoGo?," Radio I LOVE IT, last modified December 13, 2016, <http://www.radioiloveit.com/radio-production-radio-jingles-radio-imaging/abobe-audition-voco-versus-radio-journalism-ethics-sandra-muller-interview/>.

<sup>74</sup> Wesley Mattheyses and Werner Verhelst, "Audiovisual speech synthesis: An overview of the state-of-the-art," *Speech Communication* 66 (2015).

<sup>75</sup> Thies, "Face2Face."

<sup>76</sup> Pascual et al., "Language and Noise Transfer."

<sup>77</sup> Thies, Interviewed by author.

<sup>78</sup> Thies, "Face2Face." (2017) p. 5

<sup>79</sup> Lyrebird, "Partnership with the ALS association," Lyrebird, last modified April, 2018, <https://lyrebird.ai/work>.

<sup>80</sup> ALSA, "Project Revoice," The ALS Association, last modified April, 2018, <https://www.projectrevoice.org/>.

<sup>81</sup> Also known as Motor Neurone Disease - [https://en.wikipedia.org/wiki/Amyotrophic\\_lateral\\_sclerosis](https://en.wikipedia.org/wiki/Amyotrophic_lateral_sclerosis)



ability to speak. One can speculate that within several years, fully-fledged digital avatars for “self-expression” will be available to those that need them the most.

### 2.4.3 Creative uses

Finally, I would note the often-understated benefits such technologies can deliver in the creative arts and social sphere. With the introduction of rudimentary face swapping in mobile applications like Snapchat,<sup>82</sup> one envisions only a short leap to social media users uploading and sharing their faces swapped and manipulated or their voices edited for the amusement of their friends. However, given the demanding GPU requirements, time and expertise currently needed to perform even a basic deepfake, this scenario is admittedly still many years away.

In the creative arts, there is potential to expand the boundaries of what is possible for audio-visual works and other artistic pursuits. Such technology could be used to satirise, parody or critique public figures in ways that words alone could not. For my own part, I have attempted to demonstrate what is currently achievable with two compositional works based on deepfaking and text-to-speech synthesis; *Proxy*<sup>83</sup> and *Newsthink*<sup>84</sup>, although admittedly these do skirt the boundaries of unethical use. While rudimentary compared to more advanced methods of re-enactment and voice conversion, they do demonstrate the limits to which creativity can be pushed, given the inclination and necessary time investment.<sup>85</sup>

---

<sup>82</sup> Sean O’Kane, “Snapchat now lets you face swap with pictures from your camera roll,” The Verge, last modified April 22, 2016, <https://www.theverge.com/2016/4/22/11486630/snapchat-update-free-replays-face-swap-photos>.

<sup>83</sup> Nicholas Gardiner, “Proxy,” (<https://vimeo.com/278681728>; Vimeo, 2018). This piece represents a commentary on war propaganda and the interchangeable and “faceless” nature of world leaders. The work features 8 leaders that have been deepfaked based on geopolitical alliances or conflict.

<sup>84</sup> Nicholas Gardiner, “Newsthink,” (<https://vimeo.com/296263252>; Vimeo, 2018). This work is a commentary on the media manipulation, incorporating an element of controlled language and dialog as a means of creating a newsreader ‘digital avatar’. The piece employs extended applications of deepfaking, high quality text-to-speech and post video editing to effect the necessary audio and language control.

<sup>85</sup> Including unsupervised training time (i.e. leaving the computer to train on its own), *Proxy* took about 400 hours to make due to the high number of target and source actors. *Newsthink* took 100 hours, which was mostly editing the face to replace the jaw. For further info, see <https://vimeo.com/296376494>

## CHAPTER 3 – CONSEQUENCES

In this chapter, I will explore some of the potential consequences that may arise from immoral and unethical applications of this software in the public sphere.

The details of each distinct technology explored in Chapter 2 reveal an unprecedented level of creative control and innovation for the future of digital multimedia creation. The positive applications are undeniable, but their software design inherently empowers anyone to reshape the narrative of characters in a digital video. This foreshadows a frightening prospect for the future of multimedia<sup>86</sup> because, as seen in the case of deepfaking, the guiding ethical use cases and output are ultimately driven by the end-user – not the technology. From my research, three key areas of concern are apparent:

1. Social democratic discourse and information sharing
2. Personal image and avatar appropriation
3. Fraud and legal recourse

### *3.1 Social democratic discourse and information sharing*

By and large, democratic countries operate on the basis of finding popular consensus among divergent opinions. In this way, instituting policy and governance outcomes can generally be said to appropriately represent the majority view of the general public at the time. Unfortunately, this is contingent on community members maintaining a shared universe of facts and truths, supported by empirical evidence and balanced reporting. With the proliferation of social media use, the widespread democratisation of information via the internet has led to the technological amplification of unreliable, biased and duplicitous content.<sup>87</sup>

---

<sup>86</sup> CBS, "Expert warns of "terrifying" potential of digitally-altered video," Columbia Broadcasting System (CBS), last modified March 12, 2018, <https://www.cbsnews.com/news/experts-warn-of-digitally-altered-video-becoming-weaponized/>.

<sup>87</sup> Mark Verstraete and Derek E Bambauer, "Ecosystem of Distrust," *First Amendment Law Review* 16 (2017) p. 130.

In consequence, the large-scale erosion of public faith in data, statistics<sup>88</sup> and governance institutions<sup>89</sup> has meant that even the simple introduction of empirical evidence or news reporting can alienate those who have come to view statistics as elitist<sup>90</sup> and exhibit distrust towards academic and media sources.<sup>91</sup> With large sections of the public losing faith in what they see and hear, general advances in AI and machine learning capabilities will further exacerbate the devaluing of objective truth.<sup>92</sup>

With specific regard to digital avatars and social democratic discourse, the primary objective of creating fake videos would be to undermine or disparage the credibility of participants engaged in debates around truth and policy. At this point in history, video and audio are still perceived to be accurate representations of “reality” – a quality that photographs have long since lost. However, as Franklin Foer notes:

Scandalous behaviour stirs mass outrage most reliably when it is “caught on tape” ... It’s natural to trust one’s own senses, to believe what one sees – a hardwired tendency that the coming age of manipulated video will exploit... ultimately destroying faith in our strongest remaining tether to the idea of common reality.<sup>93</sup>

In terms of elected Governments and democracy, Spicer notes that technological advances in video and audio editing software, like *Face2Face*, create the potential for

---

<sup>88</sup> Williams Davis, “How statistics lost their power – and why we should fear what comes next,” The Guardian, last modified January 19, 2017, <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy>.

<sup>89</sup> Edelman, “2018 Edelman Trust Barometer - Global Report,” ([http://cms.edelman.com/sites/default/files/2018-02/2018\\_Edelman\\_Trust\\_Barometer\\_Global\\_Report\\_FEB.pdf](http://cms.edelman.com/sites/default/files/2018-02/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf); Edelman, 2018).

<sup>90</sup> Verstraete and Bambauer, “Ecosystem of Distrust.” p. 144

<sup>91</sup> Edelman, “2018 Edelman Trust Barometer Reveals Record-Breaking Drop in Trust in the U.S,” Edelman, last modified January, 2018, <https://www.edelman.com/news-awards/2018-edelman-trust-barometer-reveals-record-breaking-drop-trust-in-the-us>. In their global report, they found that the media was now the least trusted institution globally, driven primarily by a drop of trust in online platforms such as search engines and social media sites. Trust in academic sources had fallen in previous years but rose slightly in 2018.

<sup>92</sup> Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *arXiv preprint arXiv:1802.07228* (2018) p. 46.

<sup>93</sup> Franklin Foer, “The Era of Fake Video Begins,” The Atlantic, last modified May, 2018, <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>.

problematic scenarios in political campaigns.<sup>94</sup> In a divided and volatile global political landscape, the subversive effects of Russian astroturfing have already been well documented.<sup>95</sup> But given the notable worldwide escalation of propaganda and fake news from a variety of sources and vested interests in recent years,<sup>96</sup> the well-timed dispersal of video fakes into the social media sphere can potentially sway public opinions already plagued by confirmation bias<sup>97</sup>. State and non-state actors will both have a strong incentive to develop and deploy these cheaply utilised tools to influence elections both foreign and domestic to their advantage. Combined with social media bots<sup>98</sup> and other traditional forms of deceptive content such as fake news, the act of injecting false but compelling video information into a ready and willing information sharing environment can cast a shadow of illegitimacy over the entire electoral process.<sup>99</sup> For example, political consultancy firms like Cambridge Analytica could incorporate sophisticated digital forgery into their voter targeting operations,<sup>100</sup> making resultant strategic communications to recipients infinitely more convincing and manipulative.

These acts need not necessarily be confined to public or political figures. Nefarious attempts to target non-government private sector institutions and civilian movements can be used as a means of exposing social divisions and exploiting them for personal, private

---

<sup>94</sup> Robert N. Spicer, "Conclusion: Two Paths in the Legal Woods," In *Free Speech and False Speech: Political Deception and Its Legal Limits (Or Lack Thereof)* (Cham: Springer International Publishing, 2018).

<sup>95</sup> Robert S. Mueller III, "Case 1:18-cr-00032-DLF," United States Department of Justice (<https://www.justice.gov/file/1035477/download>; United States of America, 2018). For further information, see Mark Mazzetti and Katie Benner, "12 Russian Agents Indicted in Mueller Investigation," *The New York Times*, last modified July 13, 2018, <https://www.nytimes.com/2018/07/13/us/politics/mueller-indictment-russian-intelligence-hacking.html>.

<sup>96</sup> Charlie Warzel, "2017 Was The Year That The Internet Destroyed Our Shared Reality," *Buzzfeed News*, last modified December 28, 2017, <https://www.buzzfeednews.com/article/charliewarzel/2017-year-the-internet-destroyed-shared-reality>.

<sup>97</sup> Michela Del Vicario et al., "Modeling confirmation bias and polarization," *Scientific Reports* 7 (2017).

<sup>98</sup> Patrick Wang, Rafael Angarita, and Ilaria Renna, "Is this the Era of Misinformation yet? Combining Social Bots and Fake News to Deceive the Masses," (Paper presented at the Web Conference 2018 Proceedings, Lyon, France, April 23-27, 2018).

<sup>99</sup> Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* Forthcoming (2019) (2018) p. 22-23.

<sup>100</sup> Alex Hern, "Cambridge Analytica: how did it turn clicks into votes?," *The Guardian*, last modified May 6, 2018, <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>.

or geopolitical<sup>101</sup> gain. Given the partisan structure and, in some cases, highly polarised<sup>102</sup> nature of democratic systems, the ability to influence general public debate has already been well demonstrated.<sup>103</sup> In particular, community groups that evoke controversy and emotion, such as those related to abortion (e.g. Planned Parenthood), race (e.g. Black Lives Matter) or religion (e.g. Islamic congregations) would be obvious targets.

Furthermore, traditional news organisations, which usually act as a check or filter for baseless rhetoric, may abstain from reporting breaking news events for fear that the evidentiary basis supporting them will turn out to be fake.<sup>104</sup> Absent quick and reliable ways to authenticate video and audio evidence, the traditional news media risks blowback from sceptical and hyper-critical audiences in reporting factual or narrative errors, particularly forthright admission of mistakes.<sup>105</sup> Such risk will ultimately diminish the ability of the Fourth Estate to fulfil its ethical and moral obligation to spread the truth.

Finally, even the existence of digital fakes as a spectre raises the possibility of plausible deniability in cases where a real video or audio recording is released.<sup>106</sup> Vague aspersions of fakery and seeds of doubt may be enough to avoid accountability or responsibility for otherwise embarrassing evidence.<sup>107</sup>

---

<sup>101</sup> Ben Collins, Kevin Poulsen, and Spencer Ackerman, "Russia Used Facebook Events to Organize Anti-Immigrant Rallies on U.S. Soil," The Daily Beast, last modified November 9, 2017, <https://www.thedailybeast.com/exclusive-russia-used-facebook-events-to-organize-anti-immigrant-rallies-on-us-soil>.

<sup>102</sup> Verstraete and Bambauer, "Ecosystem of Distrust." p. 136. See also Md Tanvir Al Amin et al., "Crowd-sensing with polarized sources," In *2014 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, (Washington, DC: Institute of Electrical and Electronics Engineers (IEEE), 2014),

<sup>103</sup> Mueller III, "Case 1:18-cr-00032-DLF."

<sup>104</sup> Daniel Funke, "U.S. newsrooms are 'largely unprepared' to address misinformation online," Poynter, last modified November 14, 2017, <https://www.poynter.org/news/us-newsrooms-are-largely-unprepared-address-misinformation-online>.

<sup>105</sup> Verstraete and Bambauer, "Ecosystem of Distrust." p. 129

<sup>106</sup> Foer, "The Era of Fake Video".

<sup>107</sup> Maggie Haberman and Jonathan Martin, "Trump Once Said the 'Access Hollywood' Tape Was Real. Now He's Not Sure.," The New York Times, last modified November 28, 2017, <https://www.nytimes.com/2017/11/28/us/politics/trump-access-hollywood-tape.html>. If speech synthesis was a more developed field of study, Donald Trump's denial may have presented an aura of plausibility in the minds of his supporters.

### *3.2 Personal image and avatar appropriation*

As opposed to section 3.1 regarding digital avatars designed to distort social democratic discourse, personal image and avatar appropriation refers to the invasion of individual privacy and autonomy. These acts involve specifically targeted fakes designed to disparage or demean a private individual via virtual identity theft. They could include falsely generated accusations of racism, violence, inflammatory commentary, blackmail and, of course, involuntary pornography.

The deepfake phenomenon discussed in Chapter 1 has already highlighted the rise of fake celebrity pornography<sup>108</sup> which has long been feared since the early days of the internet.<sup>109</sup> However, targeted fakes made about private individuals who did not even participate in the act – pornographic or otherwise – would represent a step up in terms of debasement, illegality and invasiveness of personal autonomy. Such identity exploitation takes on a scary new life given the ease with which open-source tools can scrape photos and datasets from publicly available social media accounts, allowing end-users to quickly assemble large datasets of a target or source actor for AI training.<sup>110</sup> Furthermore, the private nature with which the exploitation is perpetrated makes the video fakes harder to debunk and more costly to discredit<sup>111</sup>, increasing the extortive value and damage they do.<sup>112</sup> Whether the objective is psychological, exploitative or financially harmful, the reputational vandalism is lasting and personalised. The ever growing number of reported

---

<sup>108</sup> Sarah Van Leuven, "Na het fake news, Fake video," *Het Nieuwsblad*, February 10, 2018.

<sup>109</sup> Jodi Dean, "Virtual Fears," *Signs: Journal of Women in Culture and Society* 24, no. 4 (1999).

<sup>110</sup> Samantha Cole, "People Are Using AI to Create Fake Porn of Their Friends and Classmates," Motherboard, last modified January 27, 2018, [https://motherboard.vice.com/en\\_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes](https://motherboard.vice.com/en_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes). In essence, the more photos and videos an individual is able to assemble of a target actor with different lighting, angles and expressions, the easier it is to assemble a composite AI model of that individual for deep learning purposes.

<sup>111</sup> Isaac Westlund, Gabriel Bianconi, and Jensen Price, "Diving Deep into Deepfakes," Medium, last modified June 3, 2016, <https://medium.com/@jensenp/diving-deep-into-deepfakes-96d2aff5f18d>. As stated in this article, it is nearly pointless to sue for harm caused by a video fake because the cost of lawyers would be greater than the damages awarded.

<sup>112</sup> Chesney and Citron, "Deep Fakes." p. 17-19

cases involving revenge pornography are one such area that video fakery would aggravate.<sup>113</sup>

Avatar appropriation can also debase the idea of human dignity, even if not being strictly illegal. Assembling deepfake, speech synthesis and re-enactment models requires the collection of large amounts of data to construct an accurate AI generated resemblance of the person involved. In cases where a target actor does not or cannot give consent, the collection of such data involves the appropriation of a human's personalised identity. Although the consequences may not be targeted to harm the individual, or illegal in nature, the acts themselves represent a betrayal of human dignity. In reference to CGI re-animation of a deceased actor, Alexi Sargeant wrote:

... it both denigrates the craft of acting and violates the dignity of the human body by treating it as a mere puppet... Leaving aside the question of consent, what would the ethical and artistic fallout be should the use of this technology become widespread.<sup>114</sup>

As stated in Chapter 2, these technologies lower the barrier to entry in avatar appropriation rather than extending its possibilities.<sup>115</sup> Using AI to resurrect, re-portray or simply replace actors and artists has been foreseen for years,<sup>116</sup> so much so that movie contracts now include clauses protecting against such practices.<sup>117</sup> But the potential could extend

---

<sup>113</sup> AAP, "'Sexual selfies' risk as one in five falls prey to revenge porn," Australian Associated Press (AAP), last modified May 8, 2017, <https://www.smh.com.au/technology/sexual-selfies-risk-as-one-in-five-fall-prey-to-revenge-porn-20170508-gvzyjb.html>. This article notes a study by The Royal Melbourne Institute of Technology and Monash University that found one in five Australians had fallen victim to revenge porn, based on a survey of 4200 people.

<sup>114</sup> Alexi Sargeant, "The Undeath of Cinema," *The New Atlantis: A Journal Of Technology & Society* (2017) p. 17.

<sup>115</sup> Hogan, Interviewed by author.

<sup>116</sup> Nick Collins, "Trading Faures: Virtual Musicians and Machine Learning," *Leonardo Music Journal*/21 (2011). Collins offers an overwrought analysis of Vocaloid's potential for virtual musicians and resynthesising deceased pop stars but was published prior to major developments such as *VoCo* and *Face2Face*.

<sup>117</sup> Reuters, "Actors Seek to Protect Posthumous Use of Big-Screen Image," *Newsweek*, last modified December 30, 2016, <https://www.newsweek.com/hollywood-actors-film-movies-carrie-fisher-537461>.

further, allowing the co-opting of said identities for commercial gain or other nefarious means, as Sargeant warns:

... practically every new technology in the world of communication and entertainment is eventually put to pornographic use, and the desires of porn users surely extend to deceased actors and celebrities. It's the next logical step after crassly commercial uses of the technology, like bringing back Audrey Hepburn to sell chocolate bars in commercials.<sup>118</sup>

### ***3.4 Fraud and legal recourse***

At this point in time, law enforcement and legal bodies may be ill-equipped to deal with the impending democratisation of video fakery that exploits and exposes human vulnerabilities.<sup>119</sup> In legal proceedings, Elizabeth Porter notes:

In recent years, images have been moving out of the evidentiary margins and are now driving argument in litigation documents from pleadings to judicial opinions. If left unregulated, visual argument threatens fundamental premises of legal discourse and decision-making.<sup>120</sup>

Evidently, their reliance on video as ostensibly accurate may undermine the legal framework on which they rely. Lawyers and courts may take language seriously, but there are no traditions in law to guide interpretation of images and no training that forces legal professionals to treat images as "entities with a complicated relationship to the real".<sup>121</sup> Paul Lester in fact predicts that:

---

<sup>118</sup> Sargeant, "The Undeath of Cinema." p. 28. While Estates and existing laws may directly prevent the use or appropriation of historic identities, one can easily imagine AI driven lookalikes being utilised for such degrading purposes.

<sup>119</sup> Brundage et al., "The Malicious Use of Artificial Intelligence." p. 6

<sup>120</sup> Elizabeth G Porter, "Taking Images Seriously," *Columbia Law Review* 114, no. 7 (2014) p. 1692.

<sup>121</sup> Rebecca Tushnet, "Worth a thousand words: the images of copyright," *Harvard Law Review* 125 (2011) p. 702.



... in a few years, images – whether still or moving – will not be allowed in trials as physical evidence because of the threat to their veracity by digital alterations... Computer technology did not start the decline in credibility of pictures, but it has hastened it.<sup>122</sup>

Likewise, anti-spoofing measures designed to protect and confirm identities may not be adequate because high-quality faked samples could be used to impersonate users and trick authentication systems.<sup>123</sup> Rita Singh, a voice forensic science expert, noted that most authentication systems – used to secure everything from banking services to smartphones – can be fooled because they rely on broad statistical features.<sup>124</sup> Furthermore, studies show that even publicly available facial and voice recognition services are vulnerable to even primitive forms of attack.<sup>125</sup> Given that researchers have developed ways to not only model and synthesise vocal phrases but also improve on the cleanliness and naturalness of low quality found data,<sup>126</sup> the threat of fraud and other forms of identity theft is a growing possibility.

Ultimately, the last few years have already laid bare the vulnerabilities of our post-truth world. The introduction of video fakes would expose them further, and appropriate recognition and development of detection mechanisms becomes a priority in combatting the consequences they would exhibit.

---

<sup>122</sup> Paul Martin Lester, "Ethical Issues & Analytical Procedures," In *Visual Ethics: A Guide for Photographers, Journalists, and Filmmakers* (Routledge: Taylor & Francis, 2018).

<sup>123</sup> Abhishek Gupta, "The Evolution Of Fraud: Ethical Implications In The Age Of Large-Scale Data Breaches And Widespread Artificial Intelligence Solutions Deployment," *ITU Journal: ICT Discoveries*, no. 1 (2018) p. 4. See also Oliver Bendel, "The synthetization of human voices," *AI & Society* 26 (2017) p. 3.

<sup>124</sup> Gent, "A.I. hears snippets of you."

<sup>125</sup> Erkam Uzun et al., "rtCaptcha: A Real-Time CAPTCHA Based Liveness Detection System," (Paper presented at the Network and Distributed Systems Security (NDSS) Symposium, San Diego, California, February 18-21, 2018).

<sup>126</sup> Lorenzo-Trueba et al., "Can we steal your vocal identity."

## CHAPTER 4 – MITIGATION

In order to evaluate the ramifications outlined in Chapter 3, it is necessary to consider how our society can mitigate the potential repercussions of their introduction into the public sphere. In this chapter, I will broadly discuss the ways in which such unethical uses of such technology can be thwarted or curtailed in their effect.

According to my research, the most effective methods can be divided into three categories:

1. Detection through technological advancements and forensic analysis
2. Regulation through legal deterrence
3. Education through social and cultural change

I will explore different methods falling under these categories and evaluate their strengths and weaknesses. However, when considering how effective each may be, it is important to contemporaneously reflect on them in the context of the current state of our democracy. Such socio and political conditions can influence and even dictate the effectiveness of each response. As such, I would note the following conclusions that were reached by academics in regard to fake news propagation and our post-truth world:

- In societies with high polarisation of views, reconstructing and agreeing on ground-truths is more difficult<sup>127</sup>
- The spread of fake news far exceeds that of truth in both speed and reach on social media sites<sup>128,129</sup>

---

<sup>127</sup> Al Amin et al., "Crowd-sensing with polarized sources."

<sup>128</sup> Soroush Vosoughi, Deb Roy, and Sinan Aral, "The spread of true and false news online," *Science* 359, no. 6380 (2018). For further information, see Seth Borenstein, "Fake News Travels Way Faster And Farther Than Truth: Study," Huffington Post, last modified March 23, 2018, [http://www.huffingtonpost.ca/2018/03/08/fake-news-study-mit\\_a\\_23380866/](http://www.huffingtonpost.ca/2018/03/08/fake-news-study-mit_a_23380866/).

<sup>129</sup> Xiaoyan Qiu et al., "Limited individual attention and online virality of low-quality information," *Nature Human Behaviour* 1, no. 7 (2017). This study found that, on social media, low quality information (not just fake news) often spreads more virally than high-quality information.

- Filter bubbles complicate the remediation of the above two<sup>130,131</sup>

#### **4.1 Detection**

A number of technological and scientific based detection methods exist today to allow the identification of video fakes. These include:

1. Multimedia forensics
2. Watermarking
3. Neural networks

##### 4.1.1 Multimedia forensics

Multimedia and image forensics techniques have been the most effective method to date for detection of faked media content. Aspects of the fields of image forensics can translate into multimedia, particularly video, which operates on similar parameters of lighting, pixel correlation and frame/image continuity. Furthermore, advances in digital forensics will lead to approaches that could automatically prove the authenticity of a clip as digital manipulation techniques improve.<sup>132</sup>

---

<sup>130</sup> Andrew Guess, Brendan Nyhan, and Jason Reifler, "Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign," In *Dartmouth Online*, (2018), <https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf>. This study found, based on consumption of fake news during the 2016 Presidential election, that filter bubbles were deep (with impacted individuals visiting on average 33 articles from fake sources) but narrow (individuals consuming fake news represented about 10% of the public).

<sup>131</sup> Eli Pariser, *The Filter Bubble: What The Internet Is Hiding From You* (London, England: Penguin UK, 2011). For further information, see Casey Newton, "The author of The Filter Bubble on how fake news is eroding trust in journalism," *The Verge*, last modified November 16, 2016, <https://www.theverge.com/2016/11/16/13653026/filter-bubble-facebook-election-eli-pariser-interview>.

<sup>132</sup> Michael Zollhöfer, "Deep Video Portraits," Stanford University, last modified 2018, [https://web.stanford.edu/~zollhofer/papers/SG2018\\_DeepVideo/page.html](https://web.stanford.edu/~zollhofer/papers/SG2018_DeepVideo/page.html).

Hany Farid is considered a leader in this field having written several articles and books on image forensics<sup>133,134</sup>, forgery detection<sup>135</sup> and exposing tampering in video<sup>136</sup>. His website, izitru<sup>137</sup>, was designed for the purpose of allowing users to upload images to determine their authenticity quickly and reliably.<sup>138</sup> However, he, along with other experts<sup>139</sup>, have noted that the introduction of AI driven digital forgery – particularly GANs<sup>140</sup> which produce faster and more accurate digital fakes – will kick off a veritable “arms race” of detection and counterfeiting.<sup>141</sup> Such adversarial models could be trained to fool detection methods by simply feeding said methods to the neural network.

More importantly as I noted at the start of this chapter, the most telling factor regarding the distribution of fake content on the internet is not so much the ability to debunk the material but how quickly and deeply it can spread. In most instances, by the time the video has been exposed as a forgery, the content would've already propagated throughout social media and internet forums. Identifying the content as fake through post-distribution forensics would not prevent the damage done. Furthermore, the existence of filter bubbles

---

<sup>133</sup> Hany Farid, *Photo forensics* (Cambridge, Massachusetts: MIT Press, 2016).

<sup>134</sup> Hany Farid, "Digital image forensics," *Scientific American* 298, no. 6 (2008).

<sup>135</sup> Hany Farid, "Image forgery detection," *IEEE Signal Processing Magazine* 26, no. 2 (2009).

<sup>136</sup> Weihong Wang and Hany Farid, "Exposing digital forgeries in interlaced and deinterlaced video," *IEEE Transactions on Information Forensics and Security* 2, no. 3 (2007).

<sup>137</sup> Fourandsix, "izitru," Fourandsix Technologies, last modified October 20, 2018, <http://fourandsix.com/>. For further information, see Rick Gladstone, "Photos Trusted But Verified," *The New York Times*, last modified May 7, 2014, <https://lens.blogs.nytimes.com/2014/05/07/photos-trusted-but-verified/>.

<sup>138</sup> This website has now been discontinued with research and development being handed over to the Defense Advanced Research Projects Agency (DARPA). See Matt Turek, "Media Forensics (MediFor)," Defense Advanced Research Projects Agency (DARPA), last modified October 20, 2018, <https://www.darpa.mil/program/media-forensics>. For further information, see Tim Mak, "Can You Believe Your Own Ears? With New 'Fake News' Tech, Not Necessarily," *National Public Radio (NPR)*, last modified April 4, 2018, <https://www.npr.org/2018/04/04/599126774/can-you-believe-your-own-ears-with-new-fake-news-tech-not-necessarily>.

<sup>139</sup> Zollhöfer, "Deep Video Portraits". Michael Zollhöfer was one of the original developers of *Face2Face*.

<sup>140</sup> Goodfellow et al., "Generative adversarial nets." For further information, see Cade Metz and Keith Collins, "How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos," *The New York Times*, last modified January 2, 2018, <https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html>.

<sup>141</sup> Will Knight, "The US military is funding an effort to catch deepfakes and other AI trickery," *MIT Technology Review*, last modified May 23, 2018, <https://www.technologyreview.com/s/611146/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/>.

means that even if content can be proven fake, said evidence may not reach all recipients or viewers of the initial forgery.

Multimedia forensics will certainly have a part to play in the detection and removal of deepfakes and other forms of digital avatar generation. However, even if the field is able to keep up with counterfeit methods of video fakery, it operates at the end of the cycle, debunking a video only once it has been created and distributed. As such, other methods of detection will be necessary to challenge the authenticity of a video before it reaches its intended audience.

#### 4.1.2 Watermarking

Watermarking allows the easy identification of manipulated digital sources via hidden signatures that indicate the editing that has taken place. Michael Zollhöfer writes in response to the development and widespread possible use of video manipulation software that:

...creative applications of video editing can and has to be flanked with continuously improved methods for forgery detection... Software companies intending to provide advanced video editing capabilities commercially could clearly watermark each video that was edited and even denote – as part of that watermark – what part and element of the scene was modified.<sup>142</sup>

Indeed, many software companies and social media sites are already looking into ways to identify digitally edited video and photo content.<sup>143</sup> Mark Randall, Vice President of Creativity at Adobe, said in relation to *VoCo*:

---

<sup>142</sup> Zollhöfer, "Deep Video Portraits".

<sup>143</sup> Danny Bradbury, "Deepfake pics and videos set off Facebook's fake news detector," Sophos, last modified September 17, 2018, <https://nakedsecurity.sophos.com/2018/09/17/deepfake-pics-and-videos-set-off-facebooks-fake-news-detector/>.

...we were already thinking of ways to add watermark technology to Project *VoCo* [and] we've gained renewed insight into the types of metadata management, watermarking or forensic technology some people desire to manage trust and authenticity in audio recording.<sup>144</sup>

Importantly, watermarks are affixed when the manipulated content is created, meaning the forgery is designated as such prior to distribution. The edited video may still be shared through social networks, but the modified elements would likely be flagged with the material, alerting recipients as such. This would avoid the pitfalls associated with end of cycle methods like multimedia forensics.

However, watermarking is not without its limitations. As has been the case with different forms of digital rights management and watermarking throughout the history of software development,<sup>145</sup> such techniques are highly likely to be broken or cracked by savvy software hackers. Hany Farid suggests that such a prospect is almost a certainty:

Any technology that will allow you to fingerprint, the adversary is going to figure out how to take it out, manipulate the content, and then put that fingerprint back in... That is almost guaranteed.<sup>146</sup>

Thus, if watermarks were relied on, they would need to be implemented concurrently with other methods.

---

<sup>144</sup> Mark Randall, "Peek Behind the Sneaks: Controversy and Opportunity in Innovation", Adobe Blog (blog), Adobe, December 12, 2016, <https://theblog.adobe.com/peek-behind-the-sneaks-controversy-and-opportunity-in-innovation/>.

<sup>145</sup> Elmar Fischer, "Denuvo says 'there is no uncrackable product'," PC Gamer, last modified August 29, 2018, <https://www.pcgamer.com/denuvo-says-there-is-no-uncrackable-product/>.

<sup>146</sup> Ariel Bogle, "'Deep fakes': How to know what's true in the fake-Obama video era," Australian Broadcasting Corporation (ABC), last modified March 4, 2018, <https://www.abc.net.au/news/science/2018-03-04/deep-fakes-and-obama-videos/9490614>.

### 4.1.3 Neural networks

Virtually all the emergent forms of video and audio manipulation discussed in this dissertation are driven by complex neural networks<sup>147</sup>. Given that current methods may not be up to the task of detecting of fake video and audio,<sup>148</sup> the logical response would be that complex problems require the application of complex solutions. In other words, utilising the same networks that generate digital forgeries to detect them.

In comparison to manual detection methods, convolutional neural networks (CNN) and other classes of neural networks operate in a machine learning context and thus exhibit more powerful image analysis features. A study by the developers of *Face2Face* using several different CNN architectures<sup>149,150</sup>, showed that their ability to detect video manipulation far exceeded that of humans.<sup>151</sup> Furthermore, other studies have shown that such networks can also be applied to detecting various signifiers of standard image forgery<sup>152</sup>, facial video forgery<sup>153</sup> and can even be used to analyse the processing history of images and videos<sup>154</sup>.

---

<sup>147</sup> Larry Hardesty, "Explained: Neural Networks," MIT News, last modified April 14, 2017, <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>.

<sup>148</sup> Samantha Cole, "Gfycat's AI Solution for Fighting Deepfakes Isn't Working," Motherboard, last modified June 19, 2018, [https://motherboard.vice.com/en\\_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn](https://motherboard.vice.com/en_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn).

<sup>149</sup> Nicolas Rahmouni et al., "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks," (Paper presented at the IEEE Workshop on Information Forensics and Security, Rennes, France, December, 2017).

<sup>150</sup> Peng Zhou et al., "Two-stream neural networks for tampered face detection," In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Honolulu, Hawaii: Institute of Electrical and Electronics Engineers (IEEE), 2017),

<sup>151</sup> Andreas Rössler et al., "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces," *arXiv preprint arXiv:1803.09179* (2018). This study found that at low video compression (low quality), humans detected the forged video only 48% of the time and were essentially reduced to guessing, whereas their XceptionNet CNN method maintained an 80-87% detection rate.

<sup>152</sup> Gregory Hill and Emily Rager, "Image Forgery Detection," In *Gregory Hill* (2018).

<sup>153</sup> Darius Afchar et al., "MesoNet: a Compact Facial Video Forgery Detection Network," *arXiv preprint arXiv:1809.00888* (2018).

<sup>154</sup> Mehdi Boroumand and Jessica Fridrich, "Deep Learning for Detecting Processing History of Images," (Paper presented at the IS&T International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics 2018, Burlingame, California, 2018).

In terms of video forgery dissemination, these AI algorithms have the potential to be deployed on information sharing platforms and embedded into social media and multimedia sites, running in the background and monitoring uploaded content for digital fakes.<sup>155</sup> Flagging, removing or alerting users to such content could potentially circumvent problems with polarisation and filter bubbles by detecting digital fakes before their widespread diffusion. Regrettably, such content would continue to propagate outside these channels, as deepfakes have continued to do so. However, the creators of fake video content would be blocked or barred from the most efficient pathways for dissemination to the wider public.

Unfortunately, as previously noted; such methods of forensic analysis often result in an “arms race” of detection and counter-detection, with each trying to gain the upper hand via improved method responses. As explained by Thies:

... it is a cat-and-mouse game, with the invention of measurements to detect fakes, one can use these measurement techniques to optimize and improve the fakes such that they cannot be detected by a certain measurement.<sup>156</sup>

GANs highlight this weakness clearly, exhibiting such powerful detection and counter-detection potential that even digital forensics experts' express scepticism at the limits of their capabilities.<sup>157</sup> Clearly, further work by researchers will be invaluable in exploring and improving their detection capabilities.

---

<sup>155</sup> Luke Dormehl, "Thanks to A.I., there is finally a way to spot 'deepfake' face swaps online," Digital Trends, last modified April 21, 2018, <https://www.digitaltrends.com/cool-tech/face-swap-recognition-algorithm/>.

<sup>156</sup> Thies, Interviewed by author.

<sup>157</sup> Knight, "The US military is funding an effort". David Gunning, the DARPA program manager, states “we don't know if there's a limit” regarding the potential of a GAN to surpass all possible detection methods and produce completely pixel accurate video fakes.



## ***4.2 Regulation***

Given the potential consequences of digital avatar creation and other forms of manipulation, it is inevitable that government and other authorities will have to respond. In his interview, Thies stressed the importance of this:

Video as proof of an evidence loses credibility. Legal authorities have to consider that change.<sup>158</sup>

### 4.2.1 Government

One immediate reaction would be for governments to outlaw the use of deepfakes and other forms of audio and video manipulation. However, there are several problems with an outright ban on digital manipulation. As Bobby Chesney and Danielle Citron explain, such a broad-brush approach could:

- stifle innovation, research and technological development
- have a chilling effect on free-speech by censoring unpopular opinions or free expression of ideas
- depending on the country, be unconstitutional
- be ultimately futile in stopping their creation, as deepfaking has shown<sup>159</sup>

Thus, authorities would have to consider a nuanced approach, curtailing the more objectionable abuses without limiting legitimate applications.

Because Photoshop and other forms of image manipulation have existed for a number of years, many of the protections from exploitation and unauthorised use of one's personal image available to individuals have already been enshrined in existing legislation and case law. These legal protections would include copyright and privacy violations, as well

---

<sup>158</sup> Thies, Interviewed by author.

<sup>159</sup> Chesney and Citron, "Deep Fakes." p. 31-33

as other civil torts. Furthermore, criminal penalties would also apply in certain more extreme circumstances. I have discussed these avenues of protection further in Section 4.2.2 below.

#### 4.2.2 Legal avenues

There are several issues with relying on existing legal protections to combat newly emergent fake multimedia:

- Due to the assembly process being divorced from real life scenarios,<sup>160</sup> the anonymity afforded to creators of deepfakes and fake content may make it difficult to identify individuals responsible
- The nature of video and audio manipulation represents a step-up from photo editing, because it is a more accurate and compelling representation of reality<sup>161</sup> and of a fully-fledged human being
- As noted previously in Chapter 3.2, acts of private or targeted harm are difficult to debunk and harder to pursue via currently available legal avenues<sup>162</sup>

Therefore, in terms of Australia's existing legal system, there are two additional options that may be worth implementation by Government and legal bodies.

---

<sup>160</sup> As noted previously, AI neural networks need a large amount of data to adequately generate deepfake, facial re-enactment and speech synthesis models. For example, *VoCo* stated that 20 minutes of speech was necessary to adequately generate an accurate vocal representation. Similarly, deepfaking requires at least several hundred images for a decent swap, but several thousand is required for better results. As noted in Chapter 3.2, such data may be scraped from user's online presence (including social media) without their consent or knowledge, thus making the perpetrator difficult to identify.

<sup>161</sup> By reference to the saying "a picture is worth a thousand words", the point made here is that a "video with audio is worth a thousand photos".

<sup>162</sup> Jessica Lake, "Watching women: Past and present legal responses to the unauthorised circulation of personal images," *Media & Arts Law Review* 21 (2016) p. 392. The author notes "civil actions require substantial resources to initiate, they often involve the victim submitting to greater publicity and scrutiny, result in damages that are difficult to recover from penniless and/or hostile defendants and offer little, if no, deterrent effect on future perpetrators".

Firstly, legislative bodies should consider a review of existing protections to ensure they are adequately representative of the changing media landscape. For example, most instances of involuntary pornography generated using a deepfake or other form of manipulation would likely already be highly illegal. Such digital fakes would fall under the purview of recently introduced amendments to the *Criminal Code Act 1995* targeting revenge pornography,<sup>163</sup> or state-based legislation<sup>164</sup>, for which the penalty is harsh<sup>165</sup>. However, the case could be made for individuals targeted by avatar forgery to seek justice through civil means as well because as Elizabeth Bird unfortunately notes, criminal avenues fall short in several areas:

- They do not allow adequate remedies for aggravated damage
- They do not allow a victim to remain anonymous
- The standards of proof<sup>166</sup> in criminal trials mean current criminal protections are not sufficient for victim redress<sup>167</sup>

In this example, it is clear the law may be lacking in avenues of reparation. A review of other existing fraud, copyright and privacy protections is also warranted.

---

<sup>163</sup> See Criminal Code Amendment (Private Sexual Material) Bill 2016 (Cth). A copy can be found at <https://www.legislation.gov.au/Details/C2016B00170>. Such material would likely fall foul of section 474.24D (a) by depicting a person who “appears to be engaged in” sexual activity, given the video is a representation of an individual via their likeness.

<sup>164</sup> Joshua Dale, “The civil implications of the crime of revenge porn,” *Privacy Law Bulletin* 14, no. 4 (2017). This article discusses current state based legislative responses that target revenge pornography in the context of evaluating possible civil action.

<sup>165</sup> AAP, “New revenge porn laws boost jail time to seven years,” Australian Associated Press (AAP), last modified August 15, 2018, <https://www.watoday.com.au/politics/federal/new-revenge-porn-laws-boost-jail-time-to-seven-years-20180815-p4zxp.html>.

<sup>166</sup> This references the burden of proof being beyond reasonable doubt in criminal cases but on the balance of probabilities in civil cases. Such a difference in standards may be crucial when combined with other complicating factors such as the potential anonymity of creators and proving who disseminated or created the fakes.

<sup>167</sup> Elizabeth Bird, “Avenging revenge porn: Why we need more than just the criminal law,” *Media & Arts Law Review* 21 (2016) p. 412-414.

Secondly, Australia should reconsider implementing a general right to be forgotten similar to the European Union<sup>168</sup>. This right allows individuals in the European Union to request that search engines like Google remove links to personal information about them. Such requests are balanced against freedom of expression and public interests.

In 2014, the Australian Law Reform Commission recommended against legislating protections similar to the right to be forgotten.<sup>169</sup> They proposed instead a privacy principle requiring deletion of personal information upon request if the individual in question provided said information.<sup>170</sup> However, this did not:

- extend to third parties – such as search engines;<sup>171</sup> and
- did not include a takedown notice scheme<sup>172</sup>

The ALRC proposal thus fell short of a broader right to be forgotten.

Under Australian law, search engines may still be liable for publishing or linking defamatory material.<sup>173</sup> Social media sites are also becoming more socially responsible when it comes to removing defamatory or illegal content,<sup>174</sup> as I note in 4.2.3 below.

However, given the speed with which fake content travels online, operators of pornography specific websites or other pseudo domains may not face the same public pressure.<sup>175</sup> Arguments have been made for takedown notice powers outside of court

---

<sup>168</sup> Mike Davis, "The right to be forgotten — international and domestic legal and policy developments," *Privacy Law Bulletin* 11, no. 8 (2014). This article outlines development of the right to be forgotten in the European Union.

<sup>169</sup> Australian Law Reform Commission, *Serious Invasions of Privacy in the Digital Era*, Discussion Paper No. 80 (2014) 6

<sup>170</sup> *Ibid.* 223 [Proposal 15-2]

<sup>171</sup> *Ibid.* 224 [Proposal 15-25]

<sup>172</sup> *Ibid.* 225 [Proposal 15-28]

<sup>173</sup> Hamish Fraser and Emma Cameron, "Why a "right to be forgotten" is not a right for Australians," *Australian Media, Technology and Communications Law Bulletin* 1, no. 3 (2014). p. 51

<sup>174</sup> *Ibid.*

<sup>175</sup> Bird, "Avenging revenge porn." p. 419

processes to minimise harm and expedite remediation,<sup>176</sup> or alternatively; a broader rights-based framework like the right to be forgotten<sup>177</sup> to address the changed multimedia landscape. Whatever the approach, these new technologies offer an unprecedented level of realistic detail while removing the need for the victim's physical participation in the defamatory act. As Bird makes clear:

Considerations of free speech and freedom of communication are negligible when it comes to revenge porn, as the public has no legitimate interest in the dissemination of the most private of private information.<sup>178</sup>

It seems that as it currently stands, the current processes may not be sufficient in protecting those targeted by video forgery. Thus, there is an obligation for Government and legal authorities to consider the legal protections available to individuals today and ask whether they are sufficient to combat the coming abundance of high-quality digital fakes.<sup>179</sup>

#### 4.2.3 Other authorities

Content platforms and social media sites which operate as purveyors of information can also ban certain types of content based upon the company's values and terms of service. Historically, they have declined to filter or block content that is legal. Furthermore, they have generally not screened for accuracy, suppression of facts, quality or opinions deemed undesirable.<sup>180</sup>

---

<sup>176</sup> *Ibid.*

<sup>177</sup> Olivia Bramwell, "A delicate balancing act: Data retention, individual privacy and the right to be forgotten in the digital age," *Media And Arts Law Review* 18, no. 2 (2013). This article argues for a broader framework of rights that encompasses the right to be forgotten but ensures a balance between privacy and freedom of expression.

<sup>178</sup> *Ibid.* p. 421

<sup>179</sup> Angela Lavoipierre and Stephen Smiley, "The nightmare of mopping up your online reputation and the 'right to be forgotten'," Australian Broadcasting Corporation (ABC), last modified July 24, 2018, <http://www.abc.net.au/news/2018-07-24/the-nightmare-of-mopping-up-your-online-reputation/10027170>.

<sup>180</sup> Verstraete and Bambauer, "Ecosystem of Distrust." p. 148-149 where they note that social media sites and search engines do not generally moderate content for truth and merely respond to terms of service and community guidelines violations.

However, with the blowback from the 2016 election and the role social media sites played in the proliferation of fake news, sites such as Facebook have begun taking a more active role,<sup>181</sup> expanding fact checking to photos and videos<sup>182</sup> and forming teams<sup>183</sup> to combat and moderate fake or misleading content. Undoubtedly, the introduction of video fakes into the public discourse would warrant similar responses.

However, outright bans similar to the one discussed in Section 4.2.1 – and even content moderation – would provoke accusations of bias and limiting free expression.<sup>184</sup> In such cases, a large measure of transparency and accountability is necessary to ensure said removal of video fakes is warranted and justified.<sup>185</sup> Such an approach may place emphasis on flagging<sup>186</sup> rather than removing such content so that users are informed but not silenced, and artistic or humorous works shared for the entertainment of others can still find an audience.<sup>187</sup>

### ***4.3 Education***

Ultimately, the nature of our interconnected world and the ease with internet users can access, view and learn the tools necessary to create digital fakes will mean that no

---

<sup>181</sup> Mark Zuckerberg, "Preparing for Elections," Facebook, last modified September 18, 2018, <https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/>.

<sup>182</sup> Antonia Woodford, "Expanding Fact-Checking to Photos and Videos," Facebook Newsroom, last modified September 13, 2018, <https://newsroom.fb.com/news/2018/09/expanding-fact-checking/>.

<sup>183</sup> Julia Boorstin, "Inside Facebook's 'war room,' where the company is fighting to stop election manipulation ahead of the midterms," CNBC, last modified October 18, 2018, <https://www.cnbc.com/2018/10/17/facebook-war-room-teams-gather-to-fight-election-manipulation.html>.

<sup>184</sup> Dawn C. Chmielewski, "Donald Trump Says Tech Companies 'Trying To Silence' Conservative Voices: 'It May Not Be Legal'," Deadline, last modified August 29, 2018, <https://deadline.com/2018/08/donald-trump-facebook-twitter-conservative-bias-claims-1202454330/>.

<sup>185</sup> Chesney and Citron, "Deep Fakes." p. 57

<sup>186</sup> Paresh Dave, "YouTube to display Wikipedia blurbs alongside conspiracy videos," Reuters, last modified March 14, 2018, <https://www.reuters.com/article/us-alphabet-youtube/youtube-to-display-wikipedia-blurbs-alongside-conspiracy-videos-idUSKCN1GP37E>.

<sup>187</sup> Carson J Reynolds, "Image Act Theory," (Paper presented at the Seventh International Conference of Computer Ethics: Philosophical Enquiry, Enschede: Centre for Telematics and Information Technology (CTIT), Netherlands, July, 2007). This article argues that an act of image manipulation is a social action that may not necessarily be designed to accuse, misrepresent or persuade but rather to entertain.

system, legal or technological, can fully stop the growing emergence of these videos. Likewise, the speed at which such technology can improve even in a few months<sup>188</sup> shows the rate at which authorities and the general public will need to respond.

The responses outlined in 4.1 and 4.2 are necessary and part of a broader effort to combat misinformation and propaganda within our society. However, the most powerful tool to stem such a tide is an educated and informed public, aware of such possibilities and critically thinking about the information they view.<sup>189,190</sup>

To this end, several media outlets have begun publicising and notifying the public of the approaching storm. Most notably, BuzzFeed issued a public warning<sup>191</sup> disguised by a deepfake and voice impersonator that was widely shared on social media.<sup>192</sup> Attempts to educate readers on how to spot a deepfake or fake video would seem less useful,<sup>193</sup> given how quickly they improve and also research suggesting that humans are sub-standard in

---

<sup>188</sup> derpfakes, "Princess Leia CGI | Deepfakes Replacement," (<https://www.youtube.com/watch?v=614we6ZaQ04>: YouTube, 2018). Comparing this early version to a remastered one, it is easy to see how quickly the technology and techniques can improve in such a short time. See derpfakes, "Princess Leia Remastered...Again | Derpfakes," (<https://www.youtube.com/watch?v=RiBqZoVe92U&feature=youtu.be>: YouTube, 2018).

<sup>189</sup> Panagiotis Metaxas, "Technology, Propaganda, and the Limits of Human Intellect," *arXiv preprint arXiv:1806.09541* (2018). In this article, it is noted that "epistemological education, recognition of self-biases and protection of our channels of communication and trusted networks are all needed to overcome the problem and continue our progress as democratic societies".

<sup>190</sup> Samantha Cole, "There Is No Tech Solution to Deepfakes," Motherboard, last modified August 15, 2018, [https://motherboard.vice.com/en\\_us/article/594qx5/there-is-no-tech-solution-to-deepfakes](https://motherboard.vice.com/en_us/article/594qx5/there-is-no-tech-solution-to-deepfakes).

<sup>191</sup> David Mack, "This PSA About Fake News From Barack Obama Is Not What It Appears " BuzzFeed News, last modified April 17, 2018, [https://www.buzzfeed.com/davidmack/obama-fake-news-jordan-peelee-psa-video-buzzfeed?utm\\_term=.sf0Q4yZ19#.ip0Y9KG8W](https://www.buzzfeed.com/davidmack/obama-fake-news-jordan-peelee-psa-video-buzzfeed?utm_term=.sf0Q4yZ19#.ip0Y9KG8W).

<sup>192</sup> BuzzFeedVideo, "You Won't Believe What Obama Says In This Video!," (<https://www.youtube.com/watch?v=cQ54GDm1eL0>: YouTube, 2018). For further information, see Michelle Castillo, "Fake video news is coming, and this clip of Obama 'insulting' Trump shows how dangerous it could be," Consumer News and Business Channel (CNBC), last modified April 18, 2018, <https://www.cnbc.com/2018/04/17/jordan-peelee-buzzfeed-psa-edits-obama-saying-things-he-never-said.html>.

<sup>193</sup> Craig Silverman, "How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video," BuzzFeed News, last modified April 17, 2018, [https://www.buzzfeed.com/craigsilverman/obama-jordan-peelee-deepfake-video-debunk-buzzfeed?utm\\_term=.xyQwxE4r1#.ibmwVW5Ll](https://www.buzzfeed.com/craigsilverman/obama-jordan-peelee-deepfake-video-debunk-buzzfeed?utm_term=.xyQwxE4r1#.ibmwVW5Ll).

detecting forged images<sup>194</sup> and video<sup>195</sup>. However, the ABC doing an educational series on these video techniques<sup>196</sup> and how deepfake videos are created<sup>197</sup> may represent a more productive approach, arming audiences with awareness, rather than tools that could soon be out of date. On this point, even developers such as Lyrebird acknowledge the importance of informing and educating the public,<sup>198</sup> conceding that audiences still fall for Photoshopped images<sup>199</sup> and other convincing fakes.<sup>200</sup>

To this end, perhaps the best encouragement for educating audiences is a need to think critically about where the video came from rather than what it purports to represent. This may mean re-establishing trust in institutions of governance and the media; a difficult task given the times.

---

<sup>194</sup> Kathryn Lynn Voith, "Recognition and Denotation of Photographic Manipulation," (PhD dissertation, Kent State University, 2017).

<sup>195</sup> Rössler et al., "FaceForensics."

<sup>196</sup> Tim Leslie, Nathan Hoad, and Ben Spraggon, "The future of fake news: Can you tell a fake video from a real one?," Australian Broadcasting Corporation, last modified September 27, 2018, <https://www.abc.net.au/news/2018-09-27/fake-news-part-one/10308638>.

<sup>197</sup> Tim Leslie, Nathan Hoad, and Ben Spraggon, "How hard is it to make a believable deepfake?," Australian Broadcasting Corporation (ABC), last modified September 28, 2018, <http://www.abc.net.au/news/2018-09-28/fake-news-how-hard-is-it-to-make-a-deepfake-video/10313906>.

<sup>198</sup> Lyrebird, "With great innovation".

<sup>199</sup> Kristina A Smith and Jung-ha Yang, "Relationship between Consumer Perceptions and Brand Preference in Photoshopped and Non-Photoshopped Online Fashion Advertisements," (Paper presented at the International Textile and Apparel Association Annual Conference Proceedings, West Virginia University, West Virginia, 2017).

<sup>200</sup> Vincent, "Lyrebird claims it can recreate any voice".



## CHAPTER 5 – ANALYSIS

In this dissertation, I have so far:

1. Detailed the research and interactions surrounding video and audio manipulation
2. Explored their intended applications
3. Hypothesised potential abuses if they were released into the public sphere; and
4. Considered the ways such abuses can be curtailed or limited

Chapter 5 will reflect on these ideas, weighing how the future of digital avatar creation will evolve and what can be done about it.

### *5.1 A fake news apocalypse?*

In early 2018, noted academic Aviv Ovadya portended that “already available tools for audio and video manipulation ... have begun to look like a potential fake news Manhattan Project”.<sup>201</sup> Indeed, reflecting on the implications enunciated in Chapter 3 as compared to and weighed against their positive benefits from Chapter 2 and mitigatory responses discussed in Chapter 4, it seems clear that there is no comparison. These technologies could pose a direct threat to our systems of governance – hardly a risk to take for minor cost savings in the multimedia industries and other assorted benefits. Even with adequate protections limiting and curtailing these threats, our divided and polarised societal conditions would ensure these tools will further drive the wedge between competing sides. The struggle for consensus would become just that much more difficult, with different groups finding “truth” in manipulated content where none may exist.

Thus, one could be forgiven for concluding that societies should act to limit and deny access to such technologies. However, the reality is not so clear cut.

---

<sup>201</sup> Charlie Warzel, “He Predicted The 2016 Fake News Crisis. Now He’s Worried About An Information Apocalypse.”, BuzzFeed News, last modified February 11, 2018, <https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news>.

## *5.2 The inevitable spread of open-source software*

It is interesting to note this quote from Randall reflecting on *VoCo*, which was posted soon after its first public demonstration:

The printing press democratised thought and drove the spread of literacy throughout the world, but it also enabled societal upheaval and religious division. The automobile allows us to travel great distances but is also a leading cause of fatal accidents. Social media connects us in meaningful ways, but it's also contributed to the spread of online bullying and "fake news" sites. Every technology comes with positive and negative consequences.<sup>202</sup>

Certainly, the progress of humankind has been marked by both our creative capacity and self-destructivity. Furthermore, in contrast to the examples noted, which overwhelmingly have progressed the development and civilisation of the human race despite their inherent drawbacks, there are many technologies that we may wish we could uninvent, such as the atomic bomb or chemical weapons. Yet one would be hard pressed to name such a time when technology development has regressed, and the creative nature of humanity halted, even for our own betterment.<sup>203</sup>

Thus, while it could be posited that these technologies may remain forever out of the public's reach, the ease with which deepfaking initially spread would render such a position naïve at best. It could also be argued that the regulatory powers of government, corporate interests or researchers<sup>204</sup> may act to limit access to them as they have to other

---

<sup>202</sup> Randall, "Peek Behind the Sneaks".

<sup>203</sup> An example from recent times may be the push to ban Lethal Autonomous Weapons. Although the systems represent a threat to human kind and an escalation of war tactics, a ban on them remains elusive. See Michael Klare, "U.S., Russia Impede Steps to Ban 'Killer Robots'," Arms Control Association, last modified October 1, 2018, <https://www.armscontrol.org/act/2018-10/news/us-russia-impede-steps-ban-%E2%80%99killer-robots%E2%80%99>.

<sup>204</sup> Westlund, Bianconi, and Price, "Diving Deep". In this study, Andreas Rossler, one of the developers of *Face2Face*, noted that they opted not to release their code publicly for fear of its harmful usage.

harmful outputs of our creative nature, such as nuclear weapons. However, as Chesney notes:

The tendency for technologies to spread only lags if they require scarce inputs that function as chokepoints to curtail access. Scarcity as a constraint on diffusion works best when the input in question is tangible and hard to obtain: plutonium or highly enriched uranium to create nuclear weapons demonstrate the point.<sup>205</sup>

As noted in Chapter 2, the input features required to access these technologies is not for want of scarcity. Demonstrations of all technologies were performed on home computers with consumer brand hardware and software. The main barrier to entry at this point in time remains the depth of knowledge in Computer Science and a requisite understanding of neural networks. However, it only takes one aspiring individual to develop and open-source these methods before it is quickly commoditised for others to use in easy self-contained applications.<sup>206</sup>

One only has to trawl the internet in the right places to know that the denial of public access to these technologies will not be possible. Following several sub-reddit's that allow safe-for-work deepfakes reveals a user advertising a cloud-based website for deepfake generation,<sup>207</sup> while a brief search of *GitHub*<sup>208</sup> quickly reveals several rudimentary facial-re-enactment projects<sup>209,210,211</sup> which demonstrate the beginnings of open-sourced avatar

---

<sup>205</sup> Chesney and Citron, "Deep Fakes." p. 8

<sup>206</sup> See Chapter 1.1 - The rise of the deepfake

<sup>207</sup> sinofis, "MachineTube," MachineTube, last modified 2018, <https://www.machine.tube/>.

<sup>208</sup> GitHub is web-based hosting service for version control, allowing users to track changes, bugs, feature requests, task management, and wikis for every project. It is mostly used for source and computer code, particularly open source or open development projects because it incentivises code-sharing and collaboration. Essentially, it is the world's largest source of open source software.

<sup>209</sup> iperov, "DeepFaceLab," GitHub, last modified August 24, 2018, <https://github.com/iperov/DeepFaceLab>.

<sup>210</sup> datitran, "face2face-demo," Github, last modified January 5, 2018, <https://github.com/datitran/face2face-demo>.

<sup>211</sup> shuvamg007, "Reenacting faces using AdversarialNets," GitHub, last modified June 8, 2018, [https://github.com/shuvamg007/facial\\_reenact\\_gan](https://github.com/shuvamg007/facial_reenact_gan).

creation<sup>212</sup>. In line with the way that deepfaking arose, their diffusion will be inevitable as user-friendly tools are developed and propagated online both in traditional and non-traditional commercial venues.<sup>213</sup> It is only a matter of time before we experience another deepfake phenomenon, albeit in a different form.

### ***5.3 A crisis of confidence***

A recent survey conducted by Stanford University students of 150 deepfake creators from *Reddit* and *4Chan* asked each about their perception of creating different deepfakes without consent. The results from this survey indicated that most felt humour was an acceptable application. However, when asked about pornographic or political content, the results were split about 50/50.<sup>214</sup> One could speculate that a result like this from an anonymous group of notoriously amoral internet users is heartening. However, it would take only a small minority of users to lay bare the full implications detailed in Chapter 3.

Indeed, at the height of deepfakes emergence, when the titled sub-reddit was still active and filled mostly with pornography, users were reflecting on the work they performed:

We are painting with revolutionary, experimental technology, one that could quite possibly shape the future of media and creative design... the safest hands for it to be might just be the general public with the power to desensitise it, rather than an exclusive few, with the power to exploit it.<sup>215</sup>

While there is merit in this argument, the acts of desensitisation and education may not best be placed in the hand of those so inclined to apply its worst designs. Indeed, as Alex Hern noted regarding this:

---

<sup>212</sup> Deepfakescovery channel, "П у т и н - в ы г о т о в ы л о в и т ь ж у л и к о в з а р у к у?," (<https://www.youtube.com/watch?v=3M0E4QnWMqA>: YouTube, 2018).

<sup>213</sup> Chesney and Citron, "Deep Fakes." p. 9

<sup>214</sup> Westlund, Bianconi, and Price, "Diving Deep".

<sup>215</sup> Hern, "My May-Thatcher deepfake".

They see themselves as harbingers of the fake news future, doing the world a service by spreading awareness of how easy it is to create convincing forgeries... That morally grey view is hard to reconcile with the way the community talks amongst itself when it thinks no one is listening. Their focus tends to be directly on the pornography.<sup>216</sup>

Thus, while it may be inevitable that these technologies will emerge one day soon and democratise video manipulation for the general public, it can be strongly argued that, in light of the deepfake phenomenon, the general public is not the best hand to guide us there.

#### ***5.4 A multi-sector response***

Having considered in Chapter 4 a broader range of approaches to combatting the unethical uses of this software, we begin to note the multi-disciplinary nature of their emergence. Although the research and development of them is heavily rooted in academe, particularly computer science & technology, their implications extend to and affect the social, cultural, legal and governmental sectors of our society. Thus, when considering the appropriate response, it is worth noting Dong Wang, who wrote:

With the greater dissemination power comes a heightened danger of misuse. It is now not only significantly easier to share ideas, but also increasingly possible to manipulate perceptions of reality at an unprecedented scale... Addressing the above challenges requires multi-disciplinary research at the intersection of computer science and social sciences.<sup>217</sup>

When asked specifically about how to combat the coming emergence of digital video forgery, Garlin Gilchrist, director of the Centre for Social Media Responsibility at the University of Michigan, suggested similarly:

---

<sup>216</sup> *Ibid.*

<sup>217</sup> Dong Wang et al., "The age of social sensing," *arXiv preprint arXiv:1801.09116* (2018).

... the metaphor is that it's an all hands-on-deck [situation]. We need the strongest thinking from every element of society: the public sector, the private sector academia, non-profits and NGOs, young people, the educational sector, different sectors of social science and computational science.<sup>218</sup>

Persuasively, the first recommendation of 26 experts<sup>219</sup> on the security implications of emerging technologies was that policymakers should collaborate closely with technical researchers to investigate, prevent and mitigate potential malicious uses of AI.<sup>220</sup> Clearly, a multi-disciplinary problem that involves technology developments and countering misinformation while protecting freedom of speech will require collaboration between stakeholders across the tech industry, journalism, government and academia.

To this end, many sectors have begun to recognise the importance of collaboration and information sharing. The Workshop on Digital Misinformation, held in conjunction with the International Conference on Web and Social Media, brought together more than 100 stakeholders from academe, media and tech companies to discuss research challenges toward a trustworthy web.<sup>221</sup> The workshop published several conclusions as follows:

- AI can play a role in defending societies from attacks in the information space. Advances in supervised and unsupervised machine learning, representation learning and natural language processing will be needed to meet this challenge.
- More AI research is needed in the study of algorithmic bias due to their:
  - o vulnerability to manipulation
  - o potential to create echo chambers; and

---

<sup>218</sup> Dawn Stover, "Garlin Gilchrist: Fighting fake news and the information apocalypse," *Bulletin of the Atomic Scientists* 74, no. 4 (2018).

<sup>219</sup> The experts were drawn from the Future of Humanity Institute, the University of Oxford, the Centre for the Study of Existential Risk, the University of Cambridge, the Centre for a New American Security, the Electronic Frontier Foundation and OpenAI.

<sup>220</sup> Brundage et al., "The Malicious Use of Artificial Intelligence."

<sup>221</sup> Giovanni Luca Ciampaglia et al., "Research Challenges of Digital Misinformation: Toward a Trustworthy Web," *AI Magazine* 39, no. 1 (2018).

- tendency to magnify cognitive and social biases.
- Reporters and fact-checking organisations need tools to help them manage the volume of digital misinformation at scale
- Support from private foundations and federal agencies will be a key ingredient for the success of future collaborative activities, the scope of which must include research, education and policy-making<sup>222</sup>

The basis of many of these recommendations when specifically applied to digital avatar creation can be found in Chapter 4 of this dissertation.

Clearly, there are looming challenges in combatting digital misinformation via video forgery. By identifying and highlighting the social, cultural, technological and governance issues, we can better address them through a multi-sector based collaborative approach that will guide these technologies, appropriately and responsibly, into the public sphere.

---

<sup>222</sup> *Ibid.* p. 72

## CHAPTER 6 – CONCLUSION

In this dissertation, I have investigated the applications and implications of four different emergent technologies being developed by different agents and at various stages of release.

In Chapter 1, I explored the emergence of deepfaking as a product of open-source software like *FakeApp* and the rise of misinformation in recent years. Noting the possibility of other similar software applications like *Face2Face*, *Adobe VoCo* and *Lyrebird* being released into the public sphere, I resolved to analyse these technologies and their implications for our society.

In Chapter 2, I detailed the research underlying each technology and their current state of development or release. I noted that the potential for end users to combine facial re-rendering with speech synthesis would allow them to generate digital avatars of any target actor in a video. I further explored their applications as intended by developers and other academics, noting their potential for positive use in the multimedia industry, creative arts and other areas of society.

In Chapter 3, I began by noting that the design of these technologies inherently means they are ethically agnostic, with their guiding ethical use ultimately being decided by the hands of those it was in. With this in mind, I hypothesised three categories of potential consequences of their misuse by nefarious actors. I focussed on three areas:

1. Social democratic discourse – the implications for information sharing in our broader society
2. Personal avatar and image appropriation – the potential for targeting of individuals by appropriating or misusing their identity
3. Fraud and legal recourse – the ability to subvert legal and identity-based protections



In Chapter 4, I analysed methods of preventing or mitigating the repercussions discussed in Chapter 3. I divided these into three potential approaches:

1. Detection via multimedia forensics, watermarking and neural networks – I concluded that methods of detection would be vital in identifying and flagging fake content but could not be relied on due to the ability to subvert detection methods and because our society remains polarised and conflicted.
2. Regulation via government, legal and other authorities – I concluded that an outright ban would not be the most efficient method for tackling video forgeries. Instead, I recommended that Government, legal bodies and other authorities review existing protections around how they engage with and protect citizens from forged video content.
3. Education – I concluded that an educated and informed public is the best protection against video fakery because it ultimately empowers citizens to moderate and filter the content themselves. However, I prefaced this by noting it would be difficult given our divided society and humans' inability to visibly detect various forms of visual fakery.

In Chapter 5, I drew together the threads of the three previous chapters and concluded that the consequences were much greater than the positive applications and potential for mitigation. Ultimately however, I found we would be unable to prevent the technology from entering into the public sphere. On this note, I concluded that the general public has a propensity to misuse these technologies whenever possible – as deepfaking has shown. Therefore, the best approach was a multi-sector collaboration of government, media, academia and privatised institutions guiding these technologies into the public hands via development and proliferation of all the methods discussed in Chapter 4.

With the rate at which new innovation in this area takes place, I would recommend further research and tracking of these emergent technologies. It is hoped that collaboration and co-operation among researchers and developers will lead to innovation and

advancement of human scientific and cultural enrichment, whilst avoiding another deepfake phenomenon.

## BIBLIOGRAPHY

- AAP. "New revenge porn laws boost jail time to seven years." Australian Associated Press (AAP), last modified August 15, 2018, <https://www.watoday.com.au/politics/federal/new-revenge-porn-laws-boost-jail-time-to-seven-years-20180815-p4zxp.html>.
- . "'Sexual selfies' risk as one in five falls prey to revenge porn." Australian Associated Press (AAP), last modified May 8, 2017, <https://www.smh.com.au/technology/sexual-selfies-risk-as-one-in-five-fall-prey-to-revenge-porn-20170508-gvzyjb.html>.
- ABC. "Cambridge Analytica harvested data from more than 87 million Facebook users, whistleblower says." Australian Broadcasting Corporation (ABC), last modified April 18, 2018, <http://www.abc.net.au/news/2018-04-18/cambridge-analytica-employee-testifies-before-uk-committee/9670192>.
- Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "MesoNet: a Compact Facial Video Forgery Detection Network." *arXiv preprint arXiv:1809.00888* (September 4, 2018).
- Al Amin, Md Tanvir, Tarek Abdelzaher, Dong Wang, and Boleslaw Szymanski. "Crowd-sensing with polarized sources." In *2014 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, (Washington, DC: Institute of Electrical and Electronics Engineers (IEEE), 2014) 67-74
- ALSA. "Project Revoice." The ALS Association, last modified April, 2018, <https://www.projectrevoice.org/>.
- Anderson, Monica, and Andrea Caumont. "How social media is reshaping news." Pew Research, last modified September 24, 2014, <http://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/>.
- Arik, Sercan O, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. "Neural Voice Cloning with a Few Samples." *arXiv preprint arXiv:1802.06006* (March 20, 2018).
- Averbuch-Elor, Hadar, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. "Bringing portraits to life." *ACM Transactions on Graphics (TOG)* 36, no. 6 (2017): 196.
- Bendel, Oliver. "The synthetization of human voices." *AI & Society* 26 (July 26, 2017): 1-7.
- Bird, Elizabeth. "Avenging revenge porn: Why we need more than just the criminal law." *Media & Arts Law Review* 21 (2016): 407-21.
- Bogle, Ariel. "'Deep fakes': How to know what's true in the fake-Obama video era." Australian Broadcasting Corporation (ABC), last modified March 4, 2018, <https://www.abc.net.au/news/science/2018-03-04/deep-fakes-and-obama-videos/9490614>.
- Boorstin, Julia. "Inside Facebook's 'war room,' where the company is fighting to stop election manipulation ahead of the midterms." CNBC, last modified October 18,

- 2018, <https://www.cnn.com/2018/10/17/facebook-war-room-teams-gather-to-fight-election-manipulation.html>.
- Borenstein, Seth. "Fake News Travels Way Faster And Farther Than Truth: Study." *Huffington Post*, last modified March 23, 2018, [http://www.huffingtonpost.ca/2018/03/08/fake-news-study-mit\\_a\\_23380866/](http://www.huffingtonpost.ca/2018/03/08/fake-news-study-mit_a_23380866/).
- Boroumand, Mehdi, and Jessica Fridrich. "Deep Learning for Detecting Processing History of Images." (Paper presented at the IS&T International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics 2018, Burlingame, California, 2018).
- Bradbury, Danny. "Deepfake pics and videos set off Facebook's fake news detector." *Sophos*, last modified September 17, 2018, <https://nakedsecurity.sophos.com/2018/09/17/deepfake-pics-and-videos-set-off-facebooks-fake-news-detector/>.
- Bramwell, Olivia. "A delicate balancing act: Data retention, individual privacy and the right to be forgotten in the digital age." *Media And Arts Law Review* 18, no. 2 (2013): 125-46.
- Lyrebird Version Beta. Lyrebird, <https://lyrebird.ai/>.
- Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, *et al.* "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv preprint arXiv:1802.07228* (2018).
- Buccafusco, Christopher, Jared Vasconcellos Grubow, and Ian Postman. "Preserving Film Preservation from the Right of Publicity." 2018 *Cardozo Law Review de novo* 1, *Cardozo Legal Studies Research Paper* No. 544(2018): Published electronically March 20, 2018. <https://ssrn.com/abstract=3145533>
- Bunker, David. "Speech2Face: Reconstructed Lip Syncing with Generative Adversarial Networks." *Data Reflexions: Thoughts and Projects*, (2017): Published electronically October 30, 2017. [https://www.dbunker.io/docs/2017\\_Bunker\\_Speech2FaceProposal.pdf](https://www.dbunker.io/docs/2017_Bunker_Speech2FaceProposal.pdf).
- BuzzFeedVideo. "You Won't Believe What Obama Says In This Video!". <https://www.youtube.com/watch?v=cQ54GDm1eL0>: YouTube, 2018.
- Castillo, Michelle. "Fake video news is coming, and this clip of Obama 'insulting' Trump shows how dangerous it could be." *Consumer News and Business Channel (CNBC)*, last modified April 18, 2018, <https://www.cnn.com/2018/04/17/jordan-peelee-buzzfeed-psa-edits-obama-saying-things-he-never-said.html>.
- CBS. "Expert warns of "terrifying" potential of digitally-altered video." *Columbia Broadcasting System (CBS)*, last modified March 12, 2018, <https://www.cbsnews.com/news/experts-warn-of-digitally-altered-video-becoming-weaponized/>.
- Chan, Caroline. "Everybody Dance Now." <https://www.youtube.com/watch?v=PCBTZh41Ris>: YouTube, 2018.

- Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. "Everybody Dance Now." *arXiv Preprint arXiv:1808.07371* (August 22, 2018).
- Chesney, Robert, and Danielle Keats Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* Forthcoming (2019) (July 14, 2018).
- Chmielewski, Dawn C. "Donald Trump Says Tech Companies "Trying To Silence" Conservative Voices: "It May Not Be Legal"." *Deadline*, last modified August 29, 2018, <https://deadline.com/2018/08/donald-trump-facebook-twitter-conservative-bias-claims-1202454330/>.
- Ciampaglia, Giovanni Luca, Alexios Mantzarlis, Gregory Maus, and Filippo Menczer. "Research Challenges of Digital Misinformation: Toward a Trustworthy Web." *AI Magazine* 39, no. 1 (2018): 65-74.
- Cloud, Adobe Creative. "#VoCo. Adobe MAX 2016 (Sneak Peeks) | Adobe Creative Cloud." <https://www.youtube.com/watch?v=I3l4XLZ59iw>: YouTube, 2016.
- Cole, Samantha. "Gfycat's AI Solution for Fighting Deepfakes Isn't Working." *Motherboard*, last modified June 19, 2018, [https://motherboard.vice.com/en\\_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn](https://motherboard.vice.com/en_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn).
- . "A.I.-Assisted Fake Porn Is Here and We're All Fucked." *Motherboard*, last modified December 11, 2017, [https://motherboard.vice.com/en\\_us/article/gydydm/gal-gadot-fake-ai-porn](https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn).
- . "People Are Using AI to Create Fake Porn of Their Friends and Classmates." *Motherboard*, last modified January 27, 2018, [https://motherboard.vice.com/en\\_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes](https://motherboard.vice.com/en_us/article/ev5eba/ai-fake-porn-of-friends-deepfakes).
- . "Reddit Just Shut Down the Deepfakes Subreddit." *Motherboard*, last modified February 7, 2018, [https://motherboard.vice.com/en\\_us/article/neqb98/reddit-shuts-down-deepfakes](https://motherboard.vice.com/en_us/article/neqb98/reddit-shuts-down-deepfakes).
- . "There Is No Tech Solution to Deepfakes." *Motherboard*, last modified August 15, 2018, [https://motherboard.vice.com/en\\_us/article/594qx5/there-is-no-tech-solution-to-deepfakes](https://motherboard.vice.com/en_us/article/594qx5/there-is-no-tech-solution-to-deepfakes).
- Collins, Ben, Kevin Poulsen, and Spencer Ackerman. "Russia Used Facebook Events to Organize Anti-Immigrant Rallies on U.S. Soil." *The Daily Beast*, last modified November 9, 2017, <https://www.thedailybeast.com/exclusive-russia-used-facebook-events-to-organize-anti-immigrant-rallies-on-us-soil>.
- Collins, Nick. "Trading Faures: Virtual Musicians and Machine Learning." *Leonardo Music Journal* 21 (2011): 35-39.
- D.N.I., Director of National Intelligence, "Background to "Assessing Russian Activities and Intentions in Recent US Elections": The Analytic Process and Cyber Incident Attribution." [https://www.dni.gov/files/documents/ICA\\_2017\\_01.pdf](https://www.dni.gov/files/documents/ICA_2017_01.pdf), Office Of The

- Director Of National Intelligence, January 6, 2017, Retrieved from <https://www.dni.gov>
- Dale, Joshua. "The civil implications of the crime of revenge porn." *Privacy Law Bulletin* 14, no. 4 (2017): 58-64.
- datitran. "face2face-demo." Github, last modified January 5, 2018, <https://github.com/datitran/face2face-demo>.
- Dave, Paresh. "YouTube to display Wikipedia blurbs alongside conspiracy videos." Reuters, last modified March 14, 2018, <https://www.reuters.com/article/us-alphabet-youtube/youtube-to-display-wikipedia-blurbs-alongside-conspiracy-videos-idUSKCN1GP37E>.
- Davis, Mike. "The right to be forgotten — international and domestic legal and policy developments." *Privacy Law Bulletin* 11, no. 8 (2014): 144-48.
- Davis, Williams. "How statistics lost their power – and why we should fear what comes next." The Guardian, last modified January 19, 2017, <https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy>.
- Dean, Jodi. "Virtual Fears." *Signs: Journal of Women in Culture and Society* 24, no. 4 (1999): 1069-78.
- FakeApp Version 2.2. deepfakeapp, <https://www.fakeapp.org/>.
- Deepfakes, Nick Cage. "Nic Cage deepfakes mini compilation." <https://www.youtube.com/watch?v=2jp4M1cIJ5A>: YouTube, 2018.
- Deepfakescovery. "Путин - вы готовы ловить жуликов за руку? ." <https://www.youtube.com/watch?v=3M0E4QnWMqA>: YouTube, 2018.
- deepfakesclub. "DeepFakes FakeApp Tutorial." DeepFakes.Club, last accessed April, 2018, <https://www.deepfakes.club/tutorial/>.
- OpenFaceSwap Version 0.9. Deepfakesclub, <https://www.deepfakes.club/openfaceswap-deepfakes-software/>.
- Del Vicario, Michela, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. "Modeling confirmation bias and polarization." *Scientific Reports* 7 (2017): 40391.
- derpfakes. "Grand Moff Tarkin | Derpfakes." <https://www.youtube.com/watch?v=6Zn1vt9vdwU&feature=youtu.be>: YouTube, 2018.
- . "Princess Leia CGI | Deepfakes Replacement." <https://www.youtube.com/watch?v=614we6ZaQ04>: YouTube, 2018.
- . "Princess Leia Remastered...Again | Derpfakes." <https://www.youtube.com/watch?v=RiBqZoVe92U&feature=youtu.be>: YouTube, 2018.

- . "Solo | A Deepfakes Story." <https://www.youtube.com/watch?v=ANXucrz7Hjs&feature=youtu.be>: YouTube, 2018.
- dhermanq. "Beta Testing #VoCo." Adobe, last modified November 8, 2016, <https://forums.adobe.com/thread/2233703>.
- Donahue, Chris, Julian McAuley, and Miller Puckette. "Synthesizing Audio with Generative Adversarial Networks." *arXiv preprint arXiv:1802.04208* (February 12, 2018).
- Dormehl, Luke. "Thanks to A.I., there is finally a way to spot 'deepfake' face swaps online." Digital Trends, last modified April 21, 2018, <https://www.digitaltrends.com/cool-tech/face-swap-recognition-algorithm/>.
- Dvorsky, George. "Deepfake Videos Are Getting Impossibly Good." Gizmodo, last modified June 14, 2018, <https://www.gizmodo.com.au/2018/06/deepfake-videos-are-getting-impossibly-good/>.
- Edelman, "2018 Edelman Trust Barometer - Global Report." [http://cms.edelman.com/sites/default/files/2018-02/2018\\_Edelman\\_Trust\\_Barometer\\_Global\\_Report\\_FEB.pdf](http://cms.edelman.com/sites/default/files/2018-02/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf), Edelman, Retrieved from <https://www.edelman.com/trust-barometer/>
- . "2018 Edelman Trust Barometer Reveals Record-Breaking Drop in Trust in the U.S." Edelman, last modified January, 2018, <https://www.edelman.com/news-awards/2018-edelman-trust-barometer-reveals-record-breaking-drop-trust-in-the-us>.
- Elor, Hadar. "Bringing Portraits to Life ". <https://www.youtube.com/watch?v=-RetOjL1Fhw>: YouTube, 2018.
- Farid, Hany. "Digital image forensics." *Scientific American* 298, no. 6 (2008): 66-71.
- . "Image forgery detection." *IEEE Signal Processing Magazine* 26, no. 2 (2009): 16-25.
- . *Photo forensics*. Cambridge, Massachusetts: MIT Press, 2016.
- Farquhar, Peter. "An AI program will soon be here to help your deepfake dancing - just don't call it deepfake." Business Insider Australia, last modified August 27, 2018, <https://www.businessinsider.com.au/artificial-intelligence-ai-deepfake-dancing-2018-8>.
- Feng, Yao, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. "Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network." *arXiv preprint arXiv:1803.07835* (March 21, 2018).
- Finkelstein, Adam. "VoCo: Text-based Insertion and Replacement in Audio Narration." <https://www.youtube.com/watch?v=RB7upq8nzlU>: YouTube, 2017.
- Fischer, Elmar. "Denuvo says 'there is no uncrackable product'." PC Gamer, last modified August 29, 2018, <https://www.pcgamer.com/denuvo-says-there-is-no-uncrackable-product/>.

- Foer, Franklin. "The Era of Fake Video Begins." *The Atlantic*, last modified May, 2018, <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>.
- Fourandsix. "izitru." Fourandsix Technologies, last modified October 20, 2018, <http://fourandsix.com/>.
- Fraser, Hamish, and Emma Cameron. "Why a "right to be forgotten" is not a right for Australians." *Australian Media, Technology and Communications Law Bulletin* 1, no. 3 (2014): 47-51.
- Funke, Daniel. "U.S. newsrooms are 'largely unprepared' to address misinformation online." Poynter, last modified November 14, 2017, <https://www.poynter.org/news/us-newsrooms-are-largely-unprepared-address-misinformation-online>.
- Fusco, Jon. "Is Adobe's Project VoCo the Photoshop for Audio?", *No Film School*, (2016): Published electronically November 4, 2016. <https://nofilmschool.com/2016/11/adobe-project-voco>.
- Gao, Yang, Rita Singh, and Bhiksha Raj. "Voice Impersonation using Generative Adversarial Networks." *arXiv preprint arXiv:1802.06840* (February 19, 2018).
- Gardiner, Nicholas. "Newsthink." <https://vimeo.com/296263252>: Vimeo, 2018.
- . "Proxy." <https://vimeo.com/278681728>: Vimeo, 2018.
- Garrido, Pablo, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track." *Computer Graphics Forum* 34, no. 2 (2015): 193-204.
- Gent, Edd. "A.I. hears snippets of you, then clones your voice." *New Scientist* 237, no. 3167 (March 3, 2018): 9.
- Giger, Thomas. "#VoCo For Radio: Revolutionary Tool Or Complete #NoGo?" *Radio I LOVE IT*, last modified December 13, 2016, <http://www.radioiloveit.com/radio-production-radio-jingles-radio-imaging/abobe-audition-voco-versus-radio-journalism-ethics-sandra-muller-interview/>.
- Gladstone, Rick. "Photos Trusted But Verified." *The New York Times*, last modified May 7, 2014, <https://lens.blogs.nytimes.com/2014/05/07/photos-trusted-but-verified/>.
- Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, (Cambridge, Massachusetts: MIT Press, 2014) 2672-80
- Guess, Andrew, Brendan Nyhan, and Jason Reifler. "Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign." *Dartmouth*, (2018): Published electronically January 9, 2018. <https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf>.



- Gupta, Abhishek. "The Evolution Of Fraud: Ethical Implications In The Age Of Large-Scale Data Breaches And Widespread Artificial Intelligence Solutions Deployment." *ITU Journal: ICT Discoveries*, no. 1 (Feb 2, 2018).
- Haberman, Maggie, and Jonathan Martin. "Trump Once Said the 'Access Hollywood' Tape Was Real. Now He's Not Sure." *The New York Times*, last modified November 28, 2017, <https://www.nytimes.com/2017/11/28/us/politics/trump-access-hollywood-tape.html>.
- Hardesty, Larry. "Explained: Neural Networks." *MIT News*, last modified April 14, 2017, <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414>.
- Hern, Alex. "Cambridge Analytica: how did it turn clicks into votes?" *The Guardian*, last modified May 6, 2018, <https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie>.
- . "My May-Thatcher deepfake won't fool you but its tech may change the world." *The Guardian*, last modified March 12, 2018, <https://www.theguardian.com/technology/2018/mar/12/may-thatcher-deepfake-face-swap-tech-change-world>.
- Hill, Gregory, and Emily Rager. "Image Forgery Detection." *Gregory Hill*, (2018): Published electronically Jun 26, 2018.
- Hogan, Josh. Interviewed by author via email correspondence. Perth, September 25, 2018.
- Hunt, Elle. "What is fake news? How to spot it and what you can do to stop it." *The Guardian*, last modified December 18, 2016, <https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>.
- iperov. "DeepFaceLab." *GitHub*, last modified August 24, 2018, <https://github.com/iperov/DeepFaceLab>.
- Jackson, Aaron S, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. "Large pose 3D face reconstruction from a single image via direct volumetric CNN regression." In *2017 IEEE International Conference on Computer Vision (ICCV)*, (Venice, Italy: Institute of Electrical and Electronics Engineers (IEEE), 2017) 1031-39
- Jalalifar, Seyed Ali, Hosein Hasani, and Hamid Aghajan. "Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks." *arXiv preprint arXiv:1803.07461* (March 20, 2018).
- Jin, Zeyu, Adam Finkelstein, Stephen DiVerdi, Jingwan Lu, and Gautham J Mysore. "CUTE: A concatenative method for voice conversion using exemplar-based unit selection." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Shanghai, China: Institute of Electrical and Electronics Engineers (IEEE), 2016) 5660-64
- Jin, Zeyu, Gautham J Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein. "VoCo: text-based insertion and replacement in audio narration." *ACM Transactions on Graphics (TOG)* 36, no. 4 (2017): 96.

- Kim, Hyeongwoo, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, *et al.* "Deep Video Portraits." *ACM Transactions on Graphics (TOG)* 37, no. 4 (2018): Article 163, 1-14.
- Klare, Michael. "U.S., Russia Impede Steps to Ban 'Killer Robots'." Arms Control Association, last modified October 1, 2018, <https://www.armscontrol.org/act/2018-10/news/us-russia-impede-steps-ban-%E2%80%98killer-robots%E2%80%99>.
- Knight, Will. "The US military is funding an effort to catch deepfakes and other AI trickery." MIT Technology Review, last modified May 23, 2018, <https://www.technologyreview.com/s/611146/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/>.
- Kumar, Rithesh. "Lyrebird - Create a digital copy of your voice." Montreal Institute for Learning Algorithms & Lyrebird, last modified September 4, 2018, <http://ritheshkumar.com/obamanet/>.
- Kumar, Rithesh, Jose Sotelo, Kundan Kumar, Alexandre de Brebisson, and Yoshua Bengio. "ObamaNet: Photo-realistic lip-sync from text." (Paper presented at the 31st Conference on Neural Information Processing Systems, Long Beach, California, December 6, 2017).
- Lake, Jessica. "Watching women: Past and present legal responses to the unauthorised circulation of personal images." *Media & Arts Law Review* 21 (2016): 383-400.
- Lavoipierre, Angela, and Stephen Smiley. "The nightmare of mopping up your online reputation and the 'right to be forgotten'." Australian Broadcasting Corporation (ABC), last modified July 24, 2018, <http://www.abc.net.au/news/2018-07-24/the-nightmare-of-mopping-up-your-online-reputation/10027170>.
- Leslie, Tim, Nathan Hoad, and Ben Spraggon. "The future of fake news: Can you tell a fake video from a real one?" Australian Broadcasting Corporation, last modified September 27, 2018, <https://www.abc.net.au/news/2018-09-27/fake-news-part-one/10308638>.
- . "How hard is it to make a believable deepfake?" Australian Broadcasting Corporation (ABC), last modified September 28, 2018, <http://www.abc.net.au/news/2018-09-28/fake-news-how-hard-is-it-to-make-a-deepfake-video/10313906>.
- Lester, Paul Martin. "Ethical Issues & Analytical Procedures." Chap. 1 In *Visual Ethics: A Guide for Photographers, Journalists, and Filmmakers*. Routledge: Taylor & Francis, 2018.
- Lorenzo-Trueba, Jaime, Fuming Fang, Xin Wang, Isao Echizen, Junichi Yamagishi, and Tomi Kinnunen. "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data." *arXiv preprint arXiv:1803.00860* (March 2, 2018).
- Lyrebird. "Lyrebird - Create a digital copy of your voice." [https://www.youtube.com/watch?v=YfU\\_sWHT8mo](https://www.youtube.com/watch?v=YfU_sWHT8mo): YouTube, 2017.

- . "Partnership with the ALS association." Lyrebird, last modified April, 2018, <https://lyrebird.ai/work>.
  - . "Vocal Avatar API." Lyrebird, last modified 2018, <https://lyrebird.ai/vocal-avatar-api>.
  - . "With great innovation comes great responsibility." In *Ethics*. <https://lyrebird.ai/ethics/>: Lyrebird, 2017.
- Mack, David. "This PSA About Fake News From Barack Obama Is Not What It Appears " BuzzFeed News, last modified April 17, 2018, [https://www.buzzfeed.com/davidmack/obama-fake-news-jordan-peepe-psa-video-buzzfeed?utm\\_term=.sf0Q4yZ19#.ip0Y9KG8W](https://www.buzzfeed.com/davidmack/obama-fake-news-jordan-peepe-psa-video-buzzfeed?utm_term=.sf0Q4yZ19#.ip0Y9KG8W).
- Mak, Tim. "Can You Believe Your Own Ears? With New 'Fake News' Tech, Not Necessarily." National Public Radio (NPR), last modified April 4, 2018, <https://www.npr.org/2018/04/04/599126774/can-you-believe-your-own-ears-with-new-fake-news-tech-not-necessarily>.
- Mattheyses, Wesley, and Werner Verhelst. "Audiovisual speech synthesis: An overview of the state-of-the-art." *Speech Communication* 66 (February 01, 2015): 182-217.
- Mazzetti, Mark, and Katie Benner. "12 Russian Agents Indicted in Mueller Investigation." The New York times, last modified July 13, 2018, <https://www.nytimes.com/2018/07/13/us/politics/mueller-indictment-russian-intelligence-hacking.html>.
- Metaxas, Panagiotis. "Technology, Propaganda, and the Limits of Human Intellect." *arXiv preprint arXiv:1806.09541* (Jun 6, 2018).
- Metz, Cade, and Keith Collins. "How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos." The New York Times, last modified January 2, 2018, <https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html>.
- Mueller III, Robert S. "Case 1:18-cr-00032-DLF." United States Department of Justice. <https://www.justice.gov/file/1035477/download>: United States of America, 2018.
- Newton, Casey. "The author of The Filter Bubble on how fake news is eroding trust in journalism." The Verge, last modified November 16, 2016, <https://www.theverge.com/2016/11/16/13653026/filter-bubble-facebook-election-eli-pariser-interview>.
- Niessner, Matthias. "Face2Face: Real-time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)." <https://www.youtube.com/watch?v=ohmajJTcpNk>: YouTube, 2016.
- O'Kane, Sean. "Snapchat now lets you face swap with pictures from your camera roll." The Verge, last modified April 22, 2016, <https://www.theverge.com/2016/4/22/11486630/snapchat-update-free-replays-face-swap-photos>.
- Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A

- generative model for raw audio." *arXiv preprint arXiv:1609.03499* (September 19, 2016).
- Papers, Two Minute. "Everybody Dance Now." <https://www.youtube.com/watch?v=cEBgi6QYDhQ>: YouTube, 2018.
- Pariser, Eli. *The Filter Bubble: What The Internet Is Hiding From You*. London, England: Penguin UK, 2011.
- Pascual, Santiago, Maruchan Park, Joan Serrà, Antonio Bonafonte, and Kang-Hun Ahn. "Language and Noise Transfer in Speech Enhancement Generative Adversarial Network." *arXiv preprint arXiv:1712.06340* (December 18, 2017).
- Pham, Hai X, Yuting Wang, and Vladimir Pavlovic. "Generative Adversarial Talking Head: Bringing Portraits to Life with a Weakly Supervised Neural Network." *arXiv preprint arXiv:1803.07716* (March 28, 2018).
- Porter, Elizabeth G. "Taking Images Seriously." *Columbia Law Review* 114, no. 7 (November, 2014): 1687-782.
- Qiu, Xiaoyan, Diego FM Oliveira, Alireza Sahami Shirazi, Alessandro Flammini, and Filippo Menczer. "Limited individual attention and online virality of low-quality information." *Nature Human Behaviour* 1, no. 7 (2017): 0132.
- Rahmouni, Nicolas, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks." (Paper presented at the IEEE Workshop on Information Forensics and Security, Rennes, France, December, 2017).
- Randall, Mark. "Peek Behind the Sneaks: Controversy and Opportunity in Innovation." In *Adobe Blog*. <https://theblog.adobe.com/peek-behind-the-sneaks-controversy-and-opportunity-in-innovation/>: Adobe, 2016.
- Reuters. "Actors Seek to Protect Posthumous Use of Big-Screen Image." *Newsweek*, last modified December 30, 2016, <https://www.newsweek.com/hollywood-actors-film-movies-carrie-fisher-537461>.
- Reynolds, Carson J. "Image Act Theory." (Paper presented at the Seventh International Conference of Computer Ethics: Philosophical Enquiry, Enschede: Centre for Telematics and Information Technology (CTIT), Netherlands, July, 2007).
- Roettgers, Janko. "Porn Producers Offer to Help Hollywood Take Down Deepfake Videos." *Variety*, last modified February 21, 2018, <https://variety.com/2018/digital/news/deepfakes-porn-adult-industry-1202705749/>.
- Rössler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces." *arXiv preprint arXiv:1803.09179* (March 24, 2018).
- Sargeant, Alexi. "The Undeath of Cinema." *The New Atlantis: A Journal Of Technology & Society* (November 17, 2017): 17-32.

- shuvamg007. "Reenacting faces using AdversarialNets." GitHub, last modified June 8, 2018, [https://github.com/shuvamg007/facial\\_reenact\\_gan](https://github.com/shuvamg007/facial_reenact_gan).
- Silverman, Craig. "How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video." BuzzFeed News, last modified April 17, 2018, [https://www.buzzfeed.com/craigsilverman/obama-jordan-peelee-deepfake-video-debunk-buzzfeed?utm\\_term=.xyQwxE4r1#.ibmwVW5LI](https://www.buzzfeed.com/craigsilverman/obama-jordan-peelee-deepfake-video-debunk-buzzfeed?utm_term=.xyQwxE4r1#.ibmwVW5LI).
- sinofis. "MachineTube." MachineTube, last modified 2018, <https://www.machine.tube/>.
- Smith, Kristina A, and Jung-ha Yang. "Relationship between Consumer Perceptions and Brand Preference in Photoshopped and Non-Photoshopped Online Fashion Advertisements." (Paper presented at the International Textile and Apparel Association Annual Conference Proceedings, West Virginia University, West Virginia, 2017).
- Sotelo, Jose, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. "Char2wav: End-to-end speech synthesis." (Paper presented at the International Conference on Learning Representations, Toulon, France, February 17, 2017).
- Spicer, Robert N. "Conclusion: Two Paths in the Legal Woods." In *Free Speech and False Speech: Political Deception and Its Legal Limits (Or Lack Thereof)*, 111-25. Cham: Springer International Publishing, 2018.
- Stover, Dawn. "Garlin Gilchrist: Fighting fake news and the information apocalypse." *Bulletin of the Atomic Scientists* 74, no. 4 (2018): 283-88.
- Suwajanakorn, Supasorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (TOG)* 36, no. 4 (2017): 95.
- Suzuki, Katsuhiko, Fumihiko Nakamura, Jiu Otsuka, Katsutoshi Masai, Yuta Itoh, Yuta Sugiura, and Maki Sugimoto. "Recognition and mapping of facial expressions to avatar by embedded photo reflective sensors in head mounted display." (Paper presented at the 2017 IEEE Virtual Reality (VR), Los Angeles, California, March 18-22, 2017).
- Theobalt, Christian. "Deep Video Portraits - SIGGRAPH 2018." <https://www.youtube.com/watch?v=qc5P2bvfl44>: YouTube, 2018.
- . "HeadOn: Real-time Reenactment of Human Portrait Videos - SIGGRAPH 2018." <https://www.youtube.com/watch?v=KRllyxqsBTM>: YouTube, 2018.
- Theobalt, Christian, Michael Zollhoefer, Marc Stamminger, Justus Thies, and Matthias Niessner. "Real-time Expression Transfer for Facial Reenactment." US Patent App. 15/256,710, 2018.
- Thies, Justus. Interviewed by author via email correspondence. Perth, September 25, 2018.
- . "Face2Face: Real-time Facial Reenactment." PhD dissertation, Friedrich Alexander University, Bavaria, 2017, European Association of Computer Graphics, 2631994.

- Thies, Justus, Michael Zollhöfer, and Matthias Nießner. "IMU2Face: Real-time Gesture-driven Facial Reenactment." *arXiv preprint arXiv:1801.01446* (December 18, 2017).
- Thies, Justus, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. "Real-time expression transfer for facial reenactment." *ACM Transactions on Graphics (TOG)* 34, no. 6 (2015): Article 183, 1-14.
- Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. "Face2Face: Real-time face capture and reenactment of RGB videos." In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (Las Vegas, Nevada: Institute of Electrical and Electronics Engineers (IEEE), 2016) 2387-95
- Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. "FaceVR: Real-Time Facial Reenactment and Eye Gaze Control in Virtual Reality." *arXiv preprint arXiv:1610.03151* (October 26, 2016).
- Thies, Justus, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. "HeadOn: Real-time Reenactment of Human Portrait Videos." *ACM Transactions on Graphics (TOG)* 37, no. 4 (2018): Article 164, 1-13.
- Timberg, Craig, and Karla Adam. "'I'm not going to be bullied by Facebook': Cambridge Analytica whistleblower tells his story." *Sydney Morning Herald*, last modified March 23, 2018, <https://www.smh.com.au/technology/i-m-not-going-to-be-bullied-by-facebook-cambridge-analytica-whistleblower-tells-his-story-20180322-p4z5sa.html>.
- Torekull, Lisa, Maria Hjorth, Tunru Julian, Haisheng Yu, and Hongting Pan. "The Future Of Storytelling And Game Writing." *Future of Media*, (2017). <http://fom.csc.kth.se/archive/files/2016-Gaming/08-Omnius-Chapter.pdf>.
- Turek, Matt. "Media Forensics (MediFor)." Defense Advanced Research Projects Agency (DARPA), last modified October 20, 2018, <https://www.darpa.mil/program/media-forensics>.
- Tushnet, Rebecca. "Worth a thousand words: the images of copyright." *Harvard Law Review* 125 (2011): 683-755.
- Uzun, Erkam, Simon Pak Ho Chung, Irfan Essa, and Wenke Lee. "rtCaptcha: A Real-Time CAPTCHA Based Liveness Detection System." (Paper presented at the Network and Distributed Systems Security (NDSS) Symposium, San Diego, California, February 18-21, 2018).
- Van Leuven, Sarah. "Na het fake news, Fake video." *Het Nieuwsblad*, February 10, 2018, 10-11.
- Verstraete, Mark, and Derek E Bambauer. "Ecosystem of Distrust." *First Amendment Law Review* 16 (March 30, 2017): 129-52.
- Vincent, James. "Lyrebird claims it can recreate any voice using just one minute of sample audio." *The Verge*, last modified April 24, 2017, <https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird>.



- . "Make a 3D model of your face from a single photo with this A.I. tool." The Verge, last modified September 18, 2017, <https://www.theverge.com/2017/9/18/16327906/3d-model-face-photograph-ai-machine-learning>.
- Voith, Kathryn Lynn. "Recognition and Denotation of Photographic Manipulation." PhD dissertation, Kent State University, Kent State, 2017, ProQuest Dissertations Publishing, 10753596.
- Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The spread of true and false news online." *Science* 359, no. 6380 (2018): 1146-51.
- Wakefield, Jane. "Social media 'outstrips T.V.' as news source for young people." British Broadcasting Corporation (BBC), last modified June 15, 2016, <http://www.bbc.com/news/uk-36528256>.
- Wang, Dong, Boleslaw K Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. "The age of social sensing." *arXiv preprint arXiv:1801.09116* (January 27, 2018).
- Wang, Patrick, Rafael Angarita, and Ilaria Renna. "Is this the Era of Misinformation yet? Combining Social Bots and Fake News to Deceive the Masses." (Paper presented at the Web Conference 2018 Proceedings, Lyon, France, April 23-27, 2018).
- Wang, Weihong, and Hany Farid. "Exposing digital forgeries in interlaced and deinterlaced video." *IEEE Transactions on Information Forensics and Security* 2, no. 3 (2007): 438-49.
- Warzel, Charlie. "2017 Was The Year That The Internet Destroyed Our Shared Reality." BuzzFeed News, last modified December 28, 2017, <https://www.buzzfeednews.com/article/charliewarzel/2017-year-the-internet-destroyed-shared-reality>.
- . "He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse." BuzzFeed News, last modified February 11, 2018, <https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news>.
- Westlund, Isaac, Gabriel Bianconi, and Jensen Price. "Diving Deep into Deepfakes." Medium, last modified June 3, 2016, <https://medium.com/@jensenp/diving-deep-into-deepfakes-96d2aff5f18d>.
- Wiki, A.I. "A Beginner's Guide to Generative Adversarial Networks (GANs)." Skymind, last modified <https://skymind.ai/wiki/generative-adversarial-network-gan>.
- Woodford, Antonia. "Expanding Fact-Checking to Photos and Videos." Facebook Newsroom, last modified September 13, 2018, <https://newsroom.fb.com/news/2018/09/expanding-fact-checking/>.
- Yee, Li Wing, and Dirk Scheiders, "Facial Landmark Tracking Final Report." Semantics Scholar, Allen Institute for Artificial Intelligence, April 17, 2017, Retrieved from <https://pdfs.semanticscholar.org/5bf0/75d3dc4a1f7d37b6e9c7fc2f97f737ccea6d4.pdf>

- Zhou, Peng, Xintong Han, Vlad I Morariu, and Larry S Davis. "Two-stream neural networks for tampered face detection." In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Honolulu, Hawaii: Institute of Electrical and Electronics Engineers (IEEE), 2017) 1831-39
- Zollhöfer, Michael. "Deep Video Portraits." Stanford University, last modified 2018, [https://web.stanford.edu/~zollhofer/papers/SG2018\\_DeepVideo/page.html](https://web.stanford.edu/~zollhofer/papers/SG2018_DeepVideo/page.html).
- Zollhöfer, Michael, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, *et al.* "Real-time non-rigid reconstruction using an RGB-D camera." *ACM Transactions on Graphics (TOG)* 33, no. 4 (2014): 156.
- Zuckerberg, Mark. "Preparing for Elections." Facebook, last modified September 18, 2018, <https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/>.



## APPENDIX A

## INTERVIEW – JUSTUS THIES

**1. What particular technology are you developing, or do you see as relevant to your field or research or work?**

Our main goal of our research projects is to give the computer the possibility to understand the world and its surrounding which is needed in a variety of applications (e.g. autonomous cars and human machine interaction).

Publications like Face2Face and HeadOn are concentrating on tracking and reconstructing the human face as well as the torso. These methods are based on the principle of analysis thru synthesis. In order to understand the surrounding, the computer tries to synthesize it. But being able to resynthesize the seen faces, enables us to modify certain properties of the face (such as expressions). Thus, the re-enactment projects are used to demonstrate the quality of our tracking approaches.

**2. What do you see as the potential beneficial applications of this technology in your field or for any other areas of society?**

In our publications we show several use cases. The most prominent field is movie editing/postproduction. It can be used to adjust the mouth movements in a dubbed movie (e.g. by transferring the movements of the dubber to the actor). We have many inquiries from international companies that want to use our technology for dubbing. Especially, advertising companies are interested.

Besides video editing, a lot of psychologists are interested in the technology. They want to analyse the effect of voice and appearance to a conversation.

**3. Do you foresee any negative applications that can arise from the use of this technology? Explain, if you can, the potential repercussions that may occur.**

We have already seen negative applications that stem from a 'democratization' of video editing tools like face swap (i.e., DeepFakes – 'FakePorn'). Also videos like the 'VaroufAKE' showed that a manipulated video is able to affect the opinion of peoples or at least to confuse them. With the upcoming democratization of video editing techniques, we will see more fake videos (similar to the history of photoshopping). Thus, video as proof of an evidence loses credibility. Legal authorities have to consider that change.

4. (DEVELOPERS ONLY) How do plan on mitigating or preparing for the implications of making this technology available to the wider public, in light of and with reference to the negative applications you outlined above?
5. (DEVELOPERS ONLY) How would you weigh the potential to mitigate the above-mentioned negative applications against the benefits these technologies can deliver to the wider public?

We are aware of our responsibility. That's why we are involved in several projects to detect fakes. We also have the goal to sensitize the society with respect to this topic. In several demonstrations and talks/interviews we speak about possible video manipulations.

FaceForensics is a recent publication that shows that we can use our manipulation methods to improve the detection of forgeries. At the moment video modifications can easily be detected (see the FaceForensics paper). But it is a cat-and-mouse game, with the invention of measurements to detect fakes, one can use these measurement techniques to optimize/improve the fakes such that they cannot be detected by a certain measurement.

Thus, I think manipulation methods will get better in the near future (also with the help of AI) and the forgery detection techniques will have a hard time to distinguish between fake and real.

## INTERVIEW – JOSHUA HOGAN

1. **What particular technology are you developing, or do you see as relevant to your field or research or work?**

Speech synthesis, CNNs, Deep Learning, GANs, Machine Learning

2. **What do you see as the potential beneficial applications of this technology in your field or for any other areas of society?**
3. **Do you foresee any negative applications that can arise from the use of this technology? Explain, if you can, the potential repercussions that may occur.**

Speech synthesis: In audio post-production, this could potentially remove the need for actors to be involved in the post-production ADR process - in theory their voices could simply be replaced synthetically, with their tone/microphone type/mic placement and room tone/reverb being emulated as well. To my mind this creates huge cost savings in terms of time and resources, but also the associated creative pitfalls - I can imagine a director with the power to replace dialogue just completely replacing the entirety of an actor's vocal performance, for example, because they started to prefer the ADR lines more. For an engineer/editor like myself, it's one of those things that adds time because of director/producer expectation of a thing that wasn't an option before - i.e. one more thing I have to do.

A good current example of this is denoising - denoising tech is getting so good now that if a director knows about it, they'll want as much of it as they can get, which adds more time to a post job. It also contributes to the creative technology problem of eroding 'destructive' creativity - what I mean by this is that if it adds another way that something can be changed after the fact (a vocal performance, for example), which changes the psychology of the original creative act/performance process - the sense that one 'never has to commit', as it can always be 'fixed' later. While I'm probably not ready to completely let go of my undo button, as a composer I'm finding a new level of depth/maturity to my work by maintaining the mindset of committing to my ideas with permanence when they happen - lots of useful techniques for this in DAW workflows, including freezing/flattening/bouncing/rendering MIDI and other DAW ideas as they happen - this is also the attraction of hardware Modular synthesis for me too - knowing that I 'can't go back' pushes me to make better decisions.

CNNs, Deep Learning, GANs, Machine Learning:

*Automated Mixing.* a neural-network/machine learning that is focussed on listening/human perception of sound could be a massive game changer and remove the need for mix engineers altogether. Plugins are already taking their first small steps in these directions. Izotope are the best mainstream example here, whose 'plugins that listen' already use a lot of the tech I think that's required. See, for example the brand-new Insight 2, which has a speech intelligibility function - it listens to dialogue, and then to the rest of the mix, and then the whole mix together and gives the dialogue intelligibility a metric/score. Also their Neutron and Ozone plugins with their mix assistants etc, as well as side-chained masking frequencies - this is already doing my job for me.

On a broader society level, I think the ramifications are much deeper than the mere creative ones that I'm contemplating above - the very nature of reality is called into question, and I think it has a polarising/siloing effect on social groups, which we're already witnessing in the age of Trump/fake news etc. It feels strange to think that I'll have to raise my 3 year old son to have a 'reality', but I think his ability to filter any mediated information that comes his way will be a key survival skill in his life. I think it will also create a more tribal/zealous society, where social groups form deeper media/viewer feedback mechanisms to confirm their own beliefs (also already happening - see social media algorithms etc.). Perhaps we are witnessing the death of mediated information altogether, and future societies will rely more on a kind of subjective approach to reality (already human nature), whereby 'objective' information/media from some external source is basically not trusted and the individual only trusts what it is that they experience subjectively.

# Re: Interview request

Justus Thies

Mon 17/09/2018 3:59 PM

To: Nick GARDINER <nfgardin@our.ecu.edu.au>;

Hi Nicholas,  
Thank you for your interest in our work.  
You can send me some questions that i will try to an  
Best,  
Justus

Am 17/09/2018 um 09:43 schrieb Nick GARDINER:

Hi Justus

I am just contacting you as a follow-up to the below request to confirm if you would still be interested in participating in my research project as a developer and expert on the subject?

I understand you would no doubt be extremely busy and have likely fielded numerous questions from the media regarding your work. Having said that, my thesis (I have attached my proposal which you can read if you like) will be a more in-depth and qualitative examination of various emergent issues. The questionnaire will likely take no more than 30 minutes to fill out.

If you are interested in participating, please let me know. Hope to hear from you.

Kind regards  
Nicholas Gardiner  
Honours Researcher  
Edith Cowan University

---

**From:** Nick GARDINER  
**Sent:** Friday, 24 August 2018 1:10 PM  
**To:**  
**Subject:** Interview request for research project

Hi Justus

My name is Nicholas Gardiner and I am an Honours student majoring in Composition & Music Technology at the Western Australian Academy of Performing Arts. I am currently writing my honours thesis on technologies like FakeApp, Lyrebird, Adobe Voco and of course; Face2Face. In essence, I was hoping I could send you some written questions regarding your work with Face2Face and subsequent developments like HeadOn.

I would say that I find the emergence of these technologies fascinating, which is why I decided to write about it. As I'm sure you are aware, there are ethical dilemmas that surround the usage of them, and whilst I aim to shed light on the development and implications of these technologies, I also wish to explore their positive potentials in industry. Ultimately, I feel that it up to us as humans to decide whether they emerge in a gradual, controlled and ethical manner (such as Face2Face and Lyrebird) or in sudden and undeniably unethical ways when open sourced in an internet free-for-all (see FakeApp and deepfaking in general).

To this end, I feel your input would be invaluable. I would love to include your perspective in my research thesis as much as I'm sure you would like your views on your own work heard. My conclusions will also explore the importance of the media and researchers such as yourself in spotlighting these developments and informing the general public.

08/11/2018

Mail - nfgardin@our.ecu.edu.au

If you are interested, I can send you through the relevant documentation and questionnaire. I hope to hear from you.

Kind regards  
Nicholas Gardiner  
Honours Researcher  
Edith Cowan University

## CONSENT FORM

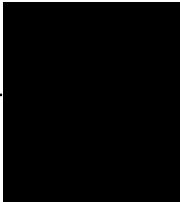
### **Facial Re-enactment, Speech Synthesis and the Rise Of The Deepfake**

- I have been provided with a letter explaining the research project and I understand the letter.
- I have been given the opportunity to ask questions and all my questions have been answered satisfactorily.
- I am aware that I can contact Dr Stuart James or the Research Ethics Officer if I have any further queries, or if I have concerns or complaints. I have been given their contact details in the Information Letter.
- I understand that participating in this project will involve completing a brief questionnaire sent via email or alternatively if requested; an online Skype interview.
- I understand that I can decline to answer any question I am uncomfortable with or unable to answer.
- I understand that I should not reveal any illegal activity that may be the subject of criminal or civil penalties in the interview.
- I understand that the researcher will be able to identify me as a particular developer/end-user but that all the information I give will be coded, kept confidential and will be accessed only by the researcher and his/her supervisor.
- I am aware that the information collected during this research will be stored in a locked cabinet at ECU for 5 years after the completion of the project and will be destroyed after that time.
- I understand that I can opt not to be identified by name or personal information in any report, thesis, or presentation of the results of this research unless I consent to being so identified. However, I will be identifiable by my role as a developer/end-user.
- I understand that I can withdraw from the research at any time without penalty.
- I freely agree to participate in this project:

NAME: Josh Hogan (Leave blank if anonymity desired)

ROLE: Interviewee

CONSENT TO BE IDENTIFIED PERSONALLY:- **YES** N

SIGNATURE:  DATE: 25/9/18