Practice Project Title: The Good, the Bad, and the Fake - Legal opinion: Proposal of a legal framework to address the threats and legal obstacles of unauthorised and malicious deepfakes

The Good, the Bad, and the Fake

Legal opinion: Proposal of a legal framework to address the threats and legal obstacles of unauthorised and malicious deepfakes[*]

Academic Year 2020/2021

[*] I would like to thank my supervisor, Professor Frederick Mostert, for his encouragement and invaluable feedback during the writing process. I also want to express my utmost gratitude to my sister, Meri Säynätkari, for her comments and help, and her unconditional support throughout my studies.

Abstract:

This legal opinion highlights both the benign and malicious uses of deepfakes and the technologies used for creating them, and illustrates how existing laws fall short in the battle against unwished forms of deepfakes. This opinion also proposes the establishment of a new global legal framework to address the regulation of deepfakes, while addressing issues from freedom of speech to legal ethics to argue that a normatively appealing and constitutionally permissible legal framework is possible.

Table of Contents

INTRODUCTION..........1

ONE – A FUCKED-UP DYSTOPIA..........3

1.1 BENIGN USE..........3
1.1.1 Education and Information..........*3*
1.1.2 Health Sector..........*5*
1.1.3 Creative Sector..........*5*
1.2 THREATS AND MALICIOUS USE..........7
1.2.1 Deepfake Pornography..........*7*
1.2.2 Mis- and Disinformation – Politics and Democracy..........*8*
1.2.3 Judicial Systems and Evidentiary Problems..........*9*
1.2.4 Corporations and Market Disturbance..........*10*
1.3 LIAR'S DIVIDEND..........12

TWO – THE EXISTING PATCHWORK..........13

2.1 COPYRIGHT..........13
2.2 RIGHT OF PUBLICITY..........14
2.3 DEFAMATION AND LIBEL..........15
2.4 PRIVACY..........15
2.5 INTENTIONAL INFLICTION OF EMOTIONAL DISTRESS (IIED)..........16
2.6 SUI GENERIS LEGISLATION..........17
2.7 DEEPFAKE DETECTION..........18

THREE – A LEGAL HEADACHE..........20

3.1 DEFINITIONS..........20
3.2 LEGAL ETHICS..........21
3.3 CROSS-BORDER NATURE OF THE INTERNET..........23
3.4 MULTIPLE STAKEHOLDERS..........24
3.4.1 Platform responsibility..........*25*
3.5 FREE SPEECH AND FREEDOM OF EXPRESSION..........26
3.6 ANONYMITY..........27
3.7 INTELLECTUAL PROPERTY RIGHTS..........28
3.8 ENFORCEMENT..........29
3.8.1 Retroactivity..........*30*
3.9 EXCEPTIONS..........31

FOUR – THE PROPOSAL..........32

4.1 THE STATUTORY TOOLBOX..........32
4.1.1 Preamble..........*32*
4.1.2 Definitions..........*32*
4.1.3 Prohibited and allowed output..........*34*
4.1.4 Scope..........*35*
4.1.5 Stakeholders and required actions..........*35*
4.1.6 Attributable rights..........*36*
4.1.7 Remedies..........*36*
4.1.8 Exceptions..........*36*
4.1.9 Establishing the organ..........*36*
4.1.10 Enforcement..........*37*
4.2 NON-LEGAL APPROACHES..........37
4.2.1 Awareness..........*37*
4.2.2 Marking..........*38*

CONCLUSION..........39

APPENDICES..........40

APPENDIX I – EXAMPLES OF DIFFERENT VISUAL DEEPFAKES..........40
APPENDIX II – USE OF DEEPFAKE TECHNOLOGIES..........42

Appendix III – Detection Tools 43
Appendix IV – Marking Methods 44

BIBLIOGRAPHY 45

Legislation 45
Cases 46
ECtHR........ 46
US 47
UK........ 47
Other........ 48
Secondary Sourers 48
Websites........ 62

Introduction

We have entered an era where we can no longer rely on what we see and hear, since deepfakes are establishing their presence on the internet. This legal opinion proposes the establishment of a new legal framework to address the regulation of deepfakes, highlights that deepfakes can have both benign and malicious uses, while illustrating how existing laws are insufficient in the battle against the unwished forms of deepfakes. Beyond enhancing awareness of the many possible uses of deepfakes and the drawbacks of the current laws, the opinion invites all relevant stakeholders to take an active role in the development of a new legal framework to tackle the modern issue. Furthermore, this legal opinion does not turn a blind eye to the issues that deepfake regulation will likely encounter, instead the opinion addresses issues from freedom of speech to legal ethics to argue that a normatively appealing and constitutionally permissible legal framework is possible.

The term 'deepfake' is a portmanteau of the words 'deep learning' and 'fake'. As the term suggests, deep learning technologies are used in creating deepfakes, and there are two main ways of generating them. One approach is to use an encoder-decoder model based on artificial intelligence (AI).[1] The more common way of the two is based on Generative Adversarial Networks (GAN).[2] The deep learning technology that deepfakes rely on trains deep neural networks, which in turn analyse large data samples to imitate facial expressions, mannerisms, voice, and inflections.[3]

For the purposes of this legal opinion, I define deepfakes as synthetic media (auditory and/or visual) generated by using deep or machine learning technologies that appear to be authentic to an ordinary person.[4] Furthermore, unauthorised deepfakes refer to the type of deepfakes that

[1] For explanation of the technology, see Jan Kietzmann and others, 'Deepfakes: Trick or Treat?' (2020) 63(2) Business Horizons 135.

[2] Ian Goodfellow and others, 'Generative Adversarial Networks' (2014) 2 Advances in Neural Information Processing Systems 2672; For more information of use of GAN for deepfakes, see Yiru Zhao and others, 'Capturing the Persistence of Facial Expression Features for Deepfake Video Detection' in Jianying Zhoud and others (eds.) *Information and Communications Security 21st International Conference, ICICS 2019* (Springer 2019) 632, 633.

[3] These data samples can also be referred to as 'training samples', 'datasets' or 'training sets'. Mika Westerlund, 'The Emergence of Deepfake Technology: A Review' (2019) 9(11) Technology Innovation Management Review 40.
See Appendix I for examples of different deepfakes.

[4] See Raina Davis, Chris Wiggins and Joan Donovan, 'Technology Factsheet: Deepfakes' in Amritha Jayanti (ed) *Tech Factsheets for Policymakers: Deepfakes* (Harvard College 2020).

use existing individuals' facial, voice, or bodily attributes to impersonate them without the individuals' consent. Additionally, deepfakes that have the ability to deceive audiences and/or to influence the audiences' perceptions of reality are categorised as malicious ones. Benign deepfakes build on the definition of deepfakes above, but do not bear malicious intentions nor create harm to individuals and/or the society as a whole.[5]

In Chapter 1, I illustrate some of the known benign and malicious uses of deepfakes. These different examples are given in order to understand the different dimensions of deepfakes, and to appreciate that deepfakes cannot be targeted with a one-size-fits-all approach. Thereafter, the current legal responses to deepfakes are analysed in Chapter 2, while highlighting their insufficiency to combat the contemporary problems. The analysis of the drawbacks of the current legal patchwork is followed by Chapter 3, which discusses the issues faced when regulating deepfakes, and how to overcome these problems. Lastly, in Chapter 4, the opinion arrives at the proposal of the legal framework, and makes suggestions on how to approach the legal drafting process.

[5] I acknowledge that using terminology of 'intent', 'threat' and 'harm' bear difficulties due to varying thresholds in different jurisdictions and possible greater burden of proof, thus using such language can be compromising.

ONE – A Fucked-up Dystopia[6]

In March 2021, the online society went crazy about magic tricks performed by Tom Cruise on TikTok. However, it was not Tom Cruise, but a deepfake created by Chris Ume, dubbed as DeepTomCruise.[7] DeepTomCruise alarmed Washington DC,[8] and got the public to talk about how authentic looking deepfakes are merely a click away. This is the part that the general public got wrong, and Ume decided to come forward as the creator in order to clarify these misconceptions. Ume told the readers that he has several years' experience in visual effects, that he collaborated with an experienced Tom Cruise impersonator (Miles Fisher) and has the equipment, trained the data sample for weeks, and used hours in post-production work.[9] So yes, deepfakes are a click away, but the good ones still require skill and labour. Hence, some caution should be reserved when discussing the quality of deepfakes.

1.1 Benign use

There is a plethora of areas for which deepfake technology could be (and has been) welcomed to benefit the society. In 2019, Citron and Chesney identified three areas in which deepfakes could be used in a beneficial manner, namely education, art, and autonomy.[10] In addition, I see that deepfakes could be realised in translations, informative content, learning, and customer service.

1.1.1 Education and Information

In the fields of education and information, deepfake technologies show great potential in, for example, documentary-making. Some film-makers have already explored the use of deepfakes

[6] Reference to BuzzFeedVideo, 'You Won't Believe What Obama Says In This Video!' (*YouTube,* 17 April 2018) <www.youtube.com/watch?v=cQ54GDm1eL0> accessed 30 March 2021.

[7] Tom, @DeepTomCruise (*TikTok*) <www.tiktok.com/@deeptomcruise?lang=en> accessed 2 July 2021.

[8] Siddharth Venkataramakrishnan, 'Behind the Tom Cruise Deepfakes That Can Evade Disinformation Tools' *Financial Times* (5 March 2021) <www.ft.com/content/721da1df-a1e5-4e2f-97fe-6de633ed4826> accessed 26 June 2021; Mark Corcoran and Matt Henry, 'The Tom Cruise Deepfake That Set Off 'Terror' in the Heart of Washington DC' *ABC News* (24 June 2021) <www.abc.net.au/news/2021-06-24/tom-cruise-deepfake-chris-ume-security-washington-dc/100234772> accessed 2 July 2021.

[9] James Vincent, 'Tom Cruise Deepfake Creator Says Public Shouldn't Be Worried About 'One-Click Fakes'' *The Verge* (5 March 2021) <www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher> accessed 26 June 2021; Emma Bowman, 'Slick Tom Cruise Deepfakes Signal That Near Flawless Forgeries May Be Here' *npr* (11 March 2021) <www.npr.org/2021/03/11/975849508/slick-tom-cruise-deepfakes-signal-that-near-flawless-forgeries-may-be-here> accessed 26 June 2021.

[10] Danielle Citron and Robert Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 CLR 1753, 1769-1771.

in their documentaries to protect the identities of victims without compromising the viewers' experience.[11] Additionally, documentaries and educational (ie teaching) materials could depict historical events and include the deepfake versions of historical characters to make learning more captivating, or alternatively depict events that would otherwise be hard to reconstruct. Regarding the reconstruction of historical events, MIT's 2019 installation *In Event of Moon Disaster* showed how deepfakes can be used to recreate historical events, even if the events portrayed in the installation did not actually take place.[12]

Although using deepfake technologies in an education or information setting is largely beneficial, we need to be cautious that the deepfake-generated materials are not considered to be authentic in the future, and that there are proper means of disclosure regarding the origins of the material. Inappropriate disclosure or lack of knowledge of the origins could lead to a falsified image of history, and, for example, the speech President Nixon delivered in MIT's installation - if detached from the original context - could in decades' time be viewed as truth.

Deepfake technology could also be used as a tool for translations. In countries with multiple official languages, or information (as well as advertising) campaigns targeted towards different peoples that do not share a language, it might become tricky to accommodate everyone if the required language skills are lacking amongst those providing the content. By using deepfake-generated translations, the target audience would receive information without compromising (too much) on the means of delivery. A good example of how to beneficially apply deepfakes in raising awareness and reaching a broad and diverse audience is David Beckham's collaboration with a Malaria awareness campaign.[13] Some companies have also started to utilise deepfake technology in their businesses, and for example, EY is using multilingual virtual doubles both in client presentations and routine communications.[14]

[11] For this specific use, see *Welcome to Chechnya* (2020).

[12] Fransesca Panetta and Halsey Burgund, *In Event of Moon Disaster* (MIT 2019) <https://moondisaster.org> accessed 18 August 2021.

[13] Malaria Must Die, 'David Beckham Speaks Nine Languages to Launch Malaria Must Die Voice Petition', (*YouTube*, 9 April 2019) < www.youtube.com/watch?v=QiiSAvKJIHo> accessed 15 March 2021.

[14] Tom Simonite, 'Deepfakes Are Now Making Business Pitches' *Wired* (16 August 2021) <www.wired.com/story/deepfakes-making-business-pitches/> accessed 16 August 2021.

1.1.2 Health Sector

Health and medical sectors are often forerunners in innovations and new technologies,[15] and exploring the possibilities of deepfakes is a natural continuum to the field's search for new ways of treatment. One area in which these sectors have started to explore the possibilities of deepfakes is in the treatment of those who have suffered from motor neurone disease, by providing new voices for those who have lost theirs.[16] Besides treatment methods, deepfake technology has been used to generate teaching materials, to improve diagnostic performance by generating deepfake radiology images.[17] Beyond human health, possibilities to use deepfake technology to improve animal welfare and using GANs to monitor animal health, and thus improve livestock farming, have been recently suggested.[18]

Not only can deepfakes and the associated technologies benefit those who suffer from physical disabilities, as they can be used in other settings within the healthcare sector. For example, as a collaboration with Unicef, MIT created deepfake images of bombed large cities (such as London and Tokyo) to enhance people's empathy towards those fleeing wars.[19] Another market for deepfakes within the healthcare sector could be therapy and care of mental health patients; some patients might rather talk with an AI-generated persona instead of with a real person, hence a deepfake generated therapist could assist people to overcome their personal burdens.[20] Furthermore, to reduce the backlog that mental health patients are experiencing, deepfake-aided therapy sessions could help to provide care in an adequate and timely manner.

1.1.3 Creative Sector

A delightful example of how to use deepfakes for good was seen in Florida's Salvador Dalí museum, when the late Spanish artist was brought to life to greet the visitors.[21] Fashion is

[15] Medical sector (both pharmaceutics and medical technology) outweighs other fields in the volume of patent applications. For example, see: 'Healthcare innovation main driver of European patent applications in 2020' (*epo.org*, 16 March 2021) <www.epo.org/news-events/news/2021/20210316.html> accessed 18 August 2021.

[16] Adam Green, 'Lawmakers and tech groups fight back against deepfakes' *Financial Times* (30 October 2019) <www.ft.com/content/b7c78624-ca57-11e9-af46-b09e8bfe60c0> accessed 29 June 2021.

[17] Vera Sorin and others, 'Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review' (2020) 27(8) Academic Radiology 1175.

[18] Suresh Neethirajan, *Beyond Deepfake Technology Fear: On its Positive Uses for Livestock Farming* (Preprints 2021) <doi: 10.20944/preprints202107.0326.v1> accessed 1 August 2021.

[19] Green (n 16).

[20] For an example of AI and deepfake related therapy, see: Deliberate <www.deliberate.ai/>.

[21] The Dalí Museum, 'Dali Lives - Art Meets Artificial Intelligence' (*YouTube*, 23 January 2019) <www.youtube.com/watch?v=Okq9zabY8rI> accessed 12 July 2021.

another creative field that has understood the possible uses of deepfakes, and has already used deepfake technology to generate fashion images.[22]

In the film industry, actors have already been performing from beyond the grave with the aid of computer-generated imagery,[23] and the industry is known to sometimes use heavy technology to deliver the storyline. Thus, it feels a rather natural transition for the entertainment industry to start employing deepfake technology when it wishes to bring deceased stars back to the screens, or de-aging actors while Hollywood turns back the clock.[24] Furthermore, as the Covid-19 pandemic has duly reminded us, transporting oneself to a certain location might be immensely demanding and sometimes hard to actualise, which might encourage the use of deepfakes for getting the job done from a distance. For example, France has already used deepfake technology to include an actress on the set when she was self-isolating.[25] In addition, we have already witnessed deepfakes entering commercials and advertising.[26]

Another possible beneficiary of deepfake technology is the gaming industry, where deepfakes could be used to improve the user experience, by creating more real-looking characters or game environments.

As illustrated, there are many possibilities to use deepfakes for good. Therefore, the proposed legal framework needs to be structured carefully so that it will not amputate the legs of a promising tool that can be of help in many sectors and scenarios.

[22] For example, Zalando has done research for using GANs to generate fashion images, Gökhan Yildirim, Calvin Seward, and Urs Bergmann, 'Disentangling Multiple Conditional Inputs in GANs' (2019) ICCV 2019 Conference Workshop on Computer Vision for Fashion, Art and Design <https://research.zalando.com/publication/generating_models_2019/> accessed 19 August 2021; Deepfakes have also been included in fashion shows: Balenciaga, 'Spring 22 Collection Note' (*balenciaga.com*) <www.balenciaga.com/en-gb/spring-22> accessed 19 August 2021.

[23] For example, Paul Walker in Fast and Furious 7: Carolyn Giardina, 'How "Furious 7" Brought the Late Paul Walker Back to Life' *The Hollywood Reporter* (11 December 2015) <www.hollywoodreporter.com/behind-screen/how-furious-7-brought-late-845763> accessed 25 January 2021; James Dean in finding Jack: Alex Ritman, 'James Dean Reborn in CGI for Vietnam War Action-Drama (Exclusive)' *The Hollywood Reporter* (6 November 2019) <www.hollywoodreporter.com/news/afm-james-dean-reborn-cgi-vietnam-war-action-drama-1252703> accessed 23 March 2021.

[24] For example, as seen in *Firefly Lane* (TV Series, 2021-), *The Irishman* (2019).

[25] Eve Jackson, 'Face swap: France's Top Soap Uses 'Deepfake' Technology for Self-Isolating Actress' *France24* (10 December 2020) <www.france24.com/en/tv-shows/encore/20201210-face-swap-france-s-top-soap-uses-deepfake-technology-for-self-isolating-actress> accessed 30 June 2021.

[26] State Farm Insurance, 'Predictions State Farm + ESPN Commercial (featuring Kenny Mayne)' (*YouTube*, 20 April 2020) <www.youtube.com/watch?v=FzOVqClci_s> accessed 30 June 2021; Hulu, 'Hulu Has Live Sports: The Deepfake Hulu Commercial' (*YouTube,* 11 September 2020) <www.youtube.com/watch?v=yPCnVeiQsUw> accessed 30 June 2021.

1.2 Threats and Malicious Use

The first publicly known use of deepfakes depicted celebrity actresses in compromised scenarios, and it is still the most common area where we encounter deepfake videos.[27] Besides synthetic pornography, deepfakes pose threats inter alia to politics and democracy, corporations, judicial systems, and (financial) markets.[28]

'Deepfakes are an agnostic and enabling technology that can be used for greater good or evil. But bad actors are often masters at weaponising technology and using it in smarter and more efficient ways than good actors.'[29] As Mostert illustrates, deepfakes are - and will be - used in all imaginable ways. Therefore, the proposed legal framework needs to be carefully drafted in a manner that captures the already known unwished uses of the technology, and those we have yet to realise.

1.2.1 Deepfake Pornography

In late 2017, Redditor deepfakes created a subreddit *r/deepfakes* where synthetic pornographic videos of female celebrities were openly shared. Deepfake pornography – maybe due to its prevailing status[30] or the fact it is the earliest use of deepfake technology[31] – has received attention from scholars and journalists regarding deepfakes. In the early days, pornographic deepfake content could be found on Pornhub, Discord, and Twitter, but these platforms have since banned these types of videos.[32] Regardless, it is not hard to find deepfake pornography, since a simple Google search provides plenty of sites that are fully dedicated to content that is synthetically generated.[33]

[27] Henry Adjer and others, *The State of Deepfakes 2019 Landscape, Threats, and Impact* (Deeptrace 2019).

[28] It should be noted that the represented examples are only few of the many possibilities.

[29] Frederick Mostert and Henry Franks, 'How to Counter Deepfakery in the Eye of the Digital Deceiver' *Financial Times* (18 June 2020) <www.ft.com/content/ea85476e-a665-11ea-92e2-cbd9b7e28ee6> accessed 12 January 2021.

[30] When Deeptrace (now Sensity) produced its first report, deepfake pornography amounted to 96% of deepfake content. Ajder and others (n 27), 1.

[31] Samantha Cole, 'AI-Assisted Fake Porn is Here and We're All Fucked' *Motherboard* (11 December 2017) <www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn> accessed 12 February 2021.

[32] Samantha Cole, 'Pornhub Is Banning AI-Generated Fake Porn Videos, Says They're Nonconsensual' *Motherboard* (6 February 2018) <www.vice.com/en/article/zmwvdw/pornhub-bans-deepfakes> accessed 31 July 2021; Samantha Cole, 'Twitter Is the Latest Platform to Ban AI-Generated Porn' *Motherboard* (7 February 2018) <www.vice.com/en/article/ywqgab/twitter-bans-deepfakes> accessed 31 July 2021.

[33] You can try to do a Google search and see how many relevant results pop up on the first page.

What makes deepfake pornography a difficult subject is its ability to invade individuals' privacy as well as sexual autonomy, and that it can be used for revenge, sexual abuse,[34] or bringing fantasies into visual context.[35] But, the existing laws covering revenge porn and distribution of non-consented sexual material might not be of aid for the depicted individuals. The obstacles these individuals face – besides the difficulty to find the person to sue – are the wordings of laws that indicate that the original picture needed to be sexual or private, and deepfake pornography usually is created by using a non-sexual image freely available online together with deepfake technology to make the final product to have a sexual tone.[36]

1.2.2 Mis- and Disinformation – Politics and Democracy

Fake news is old news. It is no longer shocking – and can even be anticipated – that some click-bait news turns out to be fake, but this is not necessarily the case with deepfakes. The two do share some similarities, such as both being prone to spreading mis- and disinformation. Disinformation aims to mislead and deceive the audience for their own nefarious purposes, whereas misinformation bears no malicious intention but is simply bad information.[37]

However, in contrast to fake news, deepfakes are in a different position because of the so-called 'processing fluency'. When psychologists use this term, they refer to our unconscious cognitive bias that favours information that our brains can process quickly – that is, audiovisual material.[38] Therefore, deepfakes are even more convincing and harmful in the current society where information found on different media outlets cannot be blindly relied on, especially since video is the most powerful communication medium in the online ecosystem.[39] Perot and Mostert have also highlighted that online users are not likely to be able to differentiate when deepfakes are used, and traditional media will also face obstacles if no relevant technology to identify deepfakes is available.[40] And, even if such technology would be acquired, it is

[34] Matt Burges, 'A Deepfake Porn Bot is Being Used to Abuse Thousands of Women' *Wired* (20 October 2020) <www.wired.co.uk/article/telegram-deepfakes-deepnude-ai> accessed 1 august 2021.

[35] Carl Öhman, 'Introducing the Pervert's Dilemma: a Contribution to the Critique of Deepfake Pornography' (2020) 22 Ethics and Information Technology 133.

[36] For example, UK Criminal Justice and Courts Act 2015 s. 35(5)(b).

[37] Nina Schick, *Deepfakes and the Infocalypse* (2020 Monoray), 11.

[38] ibid 29.

[39] ibid 11.

[40] Emma Perot and Frederick Mostert, 'Fake it till You Make it: An Examination of the US and English Approaches to Persona Protection as Applied to Deepfakes on Social Media' (2020) 15 JIPLP 32, 37.

uncertain whether traditional media would take upon the task to evaluate the contents' legitimacy.[41]

Deepfakes can be harmful in democratic discourse through disseminating disinformation. An additional burden is brought by the ever-better versions of deepfake videos, and the difficulties of citizens to detect whether the content is authentic or synthetic. This inability will eventually exhaust the citizens' critical thinking skills, further leading to a worsened ability to make informed (political) decisions and to rely on any information.[42] Especially for deepfakes in the mis- and disinformation arena, liar's dividend (1.3) is an evident obstacle, because it will increase the citizens' burden in deciding the information which to rely on.

Lastly, new areas in which mis- and disinformation can flourish are yet to be realised. One such area of disinformation that is not (yet) broadly discussed is deepfake geography, ie synthetic satellite images and/or videos. These depictions of geographical areas can be used for downplaying or covering up events, for example, where an area has been destroyed by human actions or natural disasters.[43] This illustrates that there will be new areas where false information will be distributed, and before the consumers of online content are aware of these areas and the broadness of the use of deepfakes in misrepresentation of factual information, the users' understanding of the reality is at risk.

1.2.3 Judicial Systems and Evidentiary Problems

One of the biggest threats deepfakes pose in the judicial system is evidence tampering.[44] Evidence presented in courts can be manipulated or fully produced with the aid of deepfake technology to influence the case one way or another. Moreover, problems can arise during cross-examinations, where, for example, an offering party affirms concerning details of a deepfake video when an opposing party denies the video's contents.[45] The presence of

[41] ibid.

[42] Madeleine Brady, 'Deepfakes: a New Disinformation Threat?' (*Democracy Reporting International,* August 2020) <https://democracy-reporting.org/wp-content/uploads/2020/08/2020-09-01-DRI-deepfake-publication-no-1.pdf> accessed 30 June 2021.

[43] Bo Zhao and others, 'Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence'(2021) 48(4) Cartography and Geographic Information Science 338; Hebe Campbell and Matthew Holroyd, ''Deepfake geography' could be the latest form of online disinformation' *Euronews* (7 May 2021) <www.euronews.com/2021/05/07/deepfake-geography-could-be-the-latest-form-of-online-disinformation> accessed 26 June 2021.

[44] Riana Pfefferkorn, '"Deepfakes" in the Courtroom' (2020) 29(2) B.U.Pub.Int.L.J. 245.

[45] ibid.

deepfakes might thus impact the caseloads and processing times that could increase due to deepfakes themselves. In addition, the unwelcomed presence could also increase costs and time required for authentications and verifications of the evidentiary material prior it can be admissible in courts. As an example, in a UK child custody case, the mother presented a deepfake (more precisely, a 'cheapfake')[46] audio file as evidence to the court, which eventually was proved to be artificial and dismissed by the courts.[47]

Even if the UK example shows that deepfakes can find their way before the courts, courtrooms fully banning digital evidence[48] due to the loss of confidence in their validity[49] could result in a more biased judicial system.[50] What I mean is that banning digital evidence would lend courts towards decisions that are more biased, since eyewitness testimonies have been found to be sometimes unreliable and partial.[51]

Outside of courtrooms, evidentiary issues are also finding their way to the human rights arena. Certain human rights organisations, such as WITNESS, have allocated their resources to helping citizens to use video technology as evidence for advancing human rights.[52] The human rights related work is obviously predicated on the understanding that video is seen as reliable and incorruptible evidence, and thus deepfake videos and audio are ought to cause problems in this type of evidentiary use as well.

1.2.4 Corporations and Market Disturbance

Corporations can face adverse budgetary issues due to deepfakes and deepfake technology. Deepfakes combined with social engineering attacks can be used to defraud corporations and

[46] See cheapfake definition in Britt Paris and Joan Donovan, ‘Deepfakes and Cheapfakes’ (*Data&Society* 18 September 2019) <https://datasociety.net/library/deepfakes-and-cheap-fakes/> accessed 8 August 2021

[47] Byron James ‘Why You and the Court Should not Accept Audio or Video Evidence at Face Value: How Deepfake can Be Used to Manufacture Very Plausible Evidence’ (2020) 43 IFL 41.

[48] US Federal Rules of Evidence, Rule 901(b)(9).

[49] For example, in the US, all evidence needs authentication, but ‘the bar for authentication of evidence is not particularly high.’ *United States v. Gagliardi* (2007) 506 F.3d 140, 151 (2nd Cir).

[50] Minna, 'Deepfakes: An Unknown and Unchartered Legal Landscape' (*towards data science,* 17 July 2019) <https://towardsdatascience.com/deepfakes-an-unknown-and-uncharted-legal-landscape-faec3b092eaf> accessed 29 June 2021.

[51] See, for example, Brian Cutler, Steven Penrod and Todd Martens, ‘Reliability of Eyewitness Identification the Role of system and Estimator Variables’ (1987) 11(3) Law&Hum.Behav. 233; Robert Buckhout, Daryl Figueroa, and Ethan Hoff, ‘Eyewitness identification: Effects of Suggestion and Bias in Identification from Photographs’ (1975) 6(1) Bulletin of the Psychonomic Society 71; Roderick Lindsay and Gary Wells, ‘Improving Eyewitness Identifications from Lineups: Simultaneous Versus Sequential Lineup Presentation’ (1985) 70(3) Journal of Applied Psychology 556; Thomas Albright, ‘Why eyewitness fail’ (2017) 114(30) PNAS 7758.

[52] Schick (n 37) 124-125.

thus impact inter-business negotiations and the reputation of the organisations.[53] Corporations have already witnessed such actions,[54] for example when scammers impersonated a CEO to convince the finance department employees to transfer assets to an account controlled by the scammers.[55] Yet, only a handful of corporations have plans on how to combat threats created by deepfakes.[56] Moreover, if corporations have adopted biometric technology as part of their security regime, areas in the workplace secured by such technology might be at risk if these areas were breached by using deepfakes.[57] The risks corporations are facing are not alien to state organs either,[58] and thus I emphasise the need for states to improve their national cyber security strategies to include defence tactics for deepfake scenarios.

Furthermore, in recent years we have witnessed the power of tweets and social media interactions when it comes to the financial markets. For example, in 2020 Elon Musk managed to wipe off 14 billion USD from Tesla's valuation by a single tweet,[59] and Musk's tweets are not the only concern of Tesla's shareholders. The short sellers of the electric vehicle company's stocks have been targeted by profiles that hide their identity behind deepfake generated photos, impersonating non-existing individuals.[60] It would be naïve to assume that profiles of the likes of 'Maisy Kinsley' will not flourish in the future. Instead, we need to be ever more aware of the risks that deepfakes pose to markets.

As Schick states, it is possible that more advanced profiles will use deepfakes and generate credible online footprints and deceive individuals to generate an inside-understanding of the markets and use this information to disturb the market area. And, as we have seen what type of financial impact even a single tweet can have, it is possible that deepfake technology will be

[53] Charles Owen-Jackson, 'What does the Rise of Deepfakes Mean for the Future of Cybersecurity?' (*Kapersky,* 2019) <www.kaspersky.com/blog/secure-futures-magazine/deepfakes-2019/28954/> accessed 21 June 2021.

[54] For example: Stu Sjouwerman, 'The Evolution of Deepfakes: Fighting the Next Big Threat' (*TechBeacon*) <https://techbeacon.com/security/evolution-deepfakes-fighting-next-big-threat> accessed 21 June 2021.

[55] Catherine Stupp, 'Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case', *Wall Street Journal (*30 August 2019) <www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402> accessed 21 June 2021.

[56] Kyle Wiggers, 'Fewer than 30% of Business Have Plan to Combat Deepfakes, Survey Finds' (*Venture Beat*, 24 May 2021) <https://venturebeat.com/2021/05/24/less-than-30-of-business-have-a-plan-to-combat-deepfakes-survey-finds/> accessed 22 June 2021.

[57] John Wojewidka, 'The deepfake Threat to Biometrics' (2020) 2 Biometric Technology Today 5.

[58] Jack Langa, 'Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes' (2021) 101(2) B.U.L.Rev. 761.

[59] Schick (n 37) 153.

[60] The profile of 'Maisy Kinsley' targeted Tesla's short sellers on Twitter and on LinkedIn, pressing for personal and financial information. ibid 151-153.

used to generate videos that then make falsified financial statements or 'leaked' material, which are used merely for the intention to move markets.

1.3 Liar's Dividend

The presence of deepfakes provide everyone with plausible deniability. This means that anyone can dismiss information (videos, audio, images in this case) as deepfake, if the depicted information does not serve their purposes. This phenomenon has been baptised as the liar's dividend.[61] When anything and everything can be denied, it is hard(er) to establish a reliable picture of the actual discussions and world events.

Examples of this have been already witnessed in China, Brazil, Gabon, and Malaysia, where it all culminated around whether the video material in question was real.[62] In Gabon, the uncertainty of President Bongo's video appearances sparked a failed military coup.[63] A Malaysian example shows how the legitimacy of a viral video in which two males[64] in rival political positions had sex has been debated on a national level, and there still is no certainty whether the video was authentic or a deepfake.[65] Aziz, one of the politicians depicted in the video, is still working with politics in Malaysia, and this shows that if the video was an authentic one, how Aziz enjoys the liar's dividend.

I see this phenomenon as an immediate threat for the information society in which we operate, and even Google's engineers agree with this.[66] The ability to dismiss any information will increase distrust within political and democratic discourse. It becomes an inimical cocktail when both trust and truth are decaying. When anyone or anything can be faked, everyone has plausible deniability.

[61] Citron and Chesney (n 10) 1785.

[62] Cade Metz, 'Internet Companies Prepare to Fight the 'Deepfake' Future', *The New York Times* (24 November 2019) <www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html> accessed 30 June 2021.

[63] Schick (n 37) 130-134; Sahar Cahlan, 'How Misinformation Helped Spark an Attempted Coup in Gabon', *The Washington Post* (13 February 2020) <www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/> accessed 30 June 2021.

[64] Same-sex activities are illegal in Malaysia.

[65] Schick (n 37) 135-137.

[66] Nick Dufour, one of those overseeing Google's deepfake research said that "You can already see a material effect that deepfakes have had (...)They have allowed people to claim that video evidence that would otherwise be very convincing is a fake." Metz (n 62).

TWO – The Existing Patchwork

The way we choose to perceive deepfakes is eventually reflected on the chosen discourse. I mean that if we choose to see deepfakes as weapons we will put the focus on securing identification of the use of deepfake technology, accountability, and potential remedies for those perceived as victims. If, on the other hand, deepfakes and the technology behind them are realised as tools, the focus would shift on a functional framework that would set the desired use. Nevertheless, it should be noted that deepfakes are not currently specifically addressed by civil or criminal laws,[67] and there are attempts to attack malicious deepfakes with legal tools already in existence. The question remains whether to target the technology, the deepfake as an output, or the creation and dissemination of deepfakes.

2.1 Copyright

Copyrights provide protection to the copyright holder. Deepfakes can infringe the copyright holders' moral rights[68] and their exclusive rights to the works' reproduction. In scenarios where the individual depicted in the deepfake is the copyright holder, there are some social media platform policies that provide for removals of copyright infringing material, which could aid in taking down deepfakes.[69]

However, since copyright legislation protects the author and their creations, it provides no aid to the depicted individuals if they have not created the works themselves. Thus, copyright will not be of general help to those who find themselves depicted in deepfakes without their authorisation. Moreover, copyright questions are also relevant in relation to the creation of deepfakes; who owns the copyrights to the deepfakes themselves, and if any copyrights should be granted to deepfakes at all.[70]

[67] Westerlund (n 3) 44.

[68] Especially Droit d'auteur jurisdictions, eg France, Germany.

[69] See, for example, YouTube Help, 'Submit a copyright takedown request' (*YouTube Help Centre*) <https://support.google.com/youtube/answer/2807622?hl=en-GB> accessed 18 August 2021; Instagram, 'Copyright' (*Instagram Help Centre*) <https://help.instagram.com/126382350847838> accessed 18 August 2021.

[70] WIPO, 'WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI)' (13 December 2019) WIPO/IP/AI/2/GE/20/1, available at <www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ind_lacasa.pdf> accessed 30 June.

A disparate obstacle for copyright law to be applicable to deepfakes is the fair use exception in general, but especially the transformative school of thought.[71] This outlook would justify the use of copyrighted material in creation of deepfakes since the use was transformative and thus would not infringe the original copyrights.[72]

2.2 Right of publicity

Not every country has codified law recognising image rights.[73] In the US, however, several states recognise image rights,[74] often referring to individuals' proprietary rights in their personality. These American image rights provide individuals with the possibility to prevent unauthorised use of their likeness,[75] name,[76] or personal indica (eg physical characteristics, signatures, voice).[77] Unlike with copyrights, the ability to bring a claim under image rights does not depend upon the ownership of the content.[78] It is not a given that there needs to be commercial exploitation,[79] even if commercial use is the most typical setting where appropriation of likeness occurs. As deepfakes imitate individuals' facial expressions, mannerisms, voice, and/or inflections, and thus these synthetic depictions are akin to individuals' likeness, it is probable that deepfakes could fall within the scope of the right of publicity.

Even if there are possibilities to extend image rights to deepfakes, there are two main reasons why this is not a clear-cut solution for combatting deepfakes. Firstly, image rights are not broadly nor coherently protected, and secondly, the plaintiff needs to be able to identify an

[71] On transformative use regarding fair use and copyrights, see, Pierre N. Leval, 'Towards a Fair Use Standard' (1990) 103(5) Harv.L.Rev. 1105.

[72] ibid 1111.

[73] Some EU Member States, eg France, the Netherlands, Germany, and Spain recognise image rights. See also *von Hannover v Germany (no.2)* App no 40660/08 (ECtHR, 7 February 2012), para 96.

[74] Judicially established in *Haelan Laboratories, Inc v Topps Chewing Gum, Inc,* (1953) 202 F. 2D 866 cert. denied 346 US 816, 98L. Ed. 343, 74 S. Ct. 26 (2nd Cir); California recognised in *White v Samsung Electronics America, Inc,* (1992) 971 F2d 1395.

[75] A video game depicting college athletes: *re NCAA Student-Athlete Name & Likeness Licensing Litigation* (2013) 724 F.3d 1268, 1279 (9th Cir)*;* Lookalikes: *Onassis v Christian Dior New York, Inc*, (1984) 122 Misc 2d 603, 427 NYS2d 254.

[76] *Cher v Forum International Ltd* (1982) 213 USPQ 96 (CD Cal); *Abdul-Jabbar v General Motors Corp* (1996) 85 F 3d 407.

[77] *Waits v Frito Lay, Inc,* (1992) 978 F 2d 1093 (9th Cir).

[78] Elizabeth Caldera, '"Reject the Evidence of Your Eyes and Ears": Deepfakes and the Law of Virtual Replicants' (2019) 50 Seton Hall L.Rev. 177, 192.

[79] See, for example, *Eastwood v Superior Court (National Enquirer Inc)* (1983*)* 149 Cal. App. 3d 409, in which Californian authorities established protection against non-consensual use and appropriation of plaintiff's identity when these have been to the defendant's advantage, whether commercial or otherwise.

appropriate defendant to sue them. Online platforms being a wild west of anonymity, usernames, and unverified accounts, finding someone to sue is a task close to impossible.

2.3 Defamation and Libel

In the UK, the Defamation Act 2013 and case law cover the tort of defamation.[80] The 2013 amendment introduced the serious harm threshold, and Section 1 of the Act requires that the publication 'caused or is likely to cause serious harm to the reputation' of the individual. Together with case law,[81] the notion concerning serious harm requires proof of actual harm or probable future harm, thus increasing the plaintiff's burden of proof. Furthermore, if the creator or distributor of the deepfake states that the content is a deepfake and does not actually feature the depicted individual, a defamation claim is unlikely to be successful.[82] Additionally, defamation suffers from the time limit within which the claims need to be brought. In some countries (such as Australia, England, or Scotland) the time limit is one to three years from the publication of the material,[83] and thus a late discovery of the deepfake content might make it impossible for the individual to invoke defamation claims. Nevertheless, adequate disclosures or ex-post notices of content being deepfaked do not remove the fact that even if the content is realised to be fake, it might still have consequences in the depicted individual's life.[84]

2.4 Privacy

Individuals have the right to privacy,[85] and in some jurisdictions, it can provide a cause of action against deepfakes. Privacy as an option to tackle deepfakes is most likely to occur in relation to deepfake pornography.[86] Within the European Union, the General Data Protection Regulation (GDPR)[87] and its ways to protect privacy rights can assist EU residents' battle

[80] In the US: 15 U.S.C. §1125 (Section 43(A) Lanham Act) has similar attributes as Defamation Act 2013.

[81] *Lachaux v Independent Print Ltd and another* [2019] UKSC 27.

[82] Damon Beres and Marcus Gilmer, 'A Guide to 'Deepfakes,' the Internet's Latest Moral Crisis', *Mashable* (2 February 2018), <https://mashable.com/2018/02/02/what-are-deepfakes/#pNi2cZMBtqqM> accessed 23 June 2021.

[83] Practical Law, 'Limitation Periods' (*Thomson Reuters Practical Law*, 1 May 2021) <https://uk.practicallaw.thomsonreuters.com/1-518-8770?transitionType=Default&contextData=(sc.Default)&firstPage=true> accessed 23 June 2021.

[84] For example, an Indian investigative journalist and writer spoke how the effects of deepfake pornographic video impacted her, eg through self-censoring. Rana Ayyub, 'I was the Victim of a Deepfake Porn Plot Intended to Silence Me' (*HuffPost the Blog,* 21 November 2018) <www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316> accessed 17 June 2021.

[85] In Europe, privacy rights are provided in Art 8 ECHR, and in the US in 14th Amendment (Due Process Clause) and case law (*Griswold v Connecticut* (1965) 381 U.S. 479).

[86] Danielle Citron, 'Sexual Privacy' (2019) 28 Yale L.J. 1870.

[87] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L 119.

against deepfakes, and the people of the EU can utilise GDPR namely in two ways. First, data accuracy (Article 5(1)(d)), could be used if deepfakes consist of irrelevant, inaccurate, or false content.[88] By triggering Article 17, which provides the right to be forgotten, I argue that regardless of the content being accurate or not, EU residents could request the removal of deepfake content by arguing that the depiction of them captures personal data through the voice and/or image (including audiovisual material) that is specific to the natural person.[89]

One of the problems faced with seeking protection from privacy laws stems from the fact that deepfakes are, well, fake. Deepfakes are artificially created synthetic depictions of reality, and do not actually expose any intimate details of the depicted individual's (private) life. This applies, at least as long as it is not agreed that 'the victim *becomes* the nude person in the deepfake for purposes of non-consensual pornography'.[90] Further problems evolve when storing of the deepfake's source material is considered – whether the images have been posted on private social media accounts, or made available on institutions' websites or popular blogs and public accounts of 'influencers' with millions of followers. Where the material is initially published plays a role whereupon the individual's reasonable expectations of privacy are analysed, and how the availability of the material warrants protection against deepfakes.[91]

2.5 Intentional Infliction of Emotional Distress (IIED)

The common law tort of intentional infliction of emotional distress allows individuals to recover for emotional distress that is caused intentionally by another individual.[92] Some deepfake pornography films have left the depicted individuals in 'so much pain',[93] being 'incredibly anxious about even leaving the house',[94] and silencing them.[95] These experiences of deepfake pornography victims could be classified under emotional distress resulting from the deepfakes being created and distributed.

[88] Betül Çolak, 'Legal Issues of Deepfakes' (*Institute for Internet & the Just Society*, 19 January 2021) <www.internetjustsociety.org/legal-issues-of-deepfakes> accessed 29 June 2021.

[89] In accordance with the definition of personal data in GDPR Art. 4(1).

[90] Douglas Harris, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You' (2019) 7 Duke Law and Technology Review 99, 123.

[91] Kelsey Farish, 'Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California's Publicity Right in the Age of the Deepfake' (2020) 15 JIPLP 40, 45.

[92] Erik Gerstner, 'Face/Off: "DeepFake" Face Swaps and Privacy Laws' (2020) 87(1) D.C.J. 1.

[93] Daniella Scott, 'Deepfake Porn Nearly Ruined My Life' *Elle UK* (6 February 2020) <www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/> accessed 9 July 2021.

[94] Sara Royle, ''Deepfake porn images still give me nightmares'' *BBC* (6 January 2021) <www.bbc.com/news/technology-55546372> accessed 4 March 2021.

[95] Schick (n 37) 127-128.

Unlike a defamation claim that has the requirement of false statement, an IIED claim can be used by the victims without the need to analyse the authenticity of the video's content. Not only IIED claims face the same difficulties of finding a person to sue, but on top of that the element of intent needs to be fulfilled. The intent to cause emotional distress might be even harder if the creator is found to make this type of content for mere entertainment. Further uncertainties regarding the use of IIED claims concern public figures, as the plaintiff needs to prove that the statement causing emotional distress has an intent that the receiving audience believe the statement to be true.[96]

2.6 Sui Generis Legislation

During the past few years, only a few states have taken action in regulating deepfakes through new bills and amendments to their existing laws. Regardless of the lack of codification in this realm, the need (and willingness) for legislating deepfakes is well acknowledged.[97]

In the US, legislative action concerning deepfakes has emerged both on federal and state level. Although the majority of the federal laws focus on researching[98] and reporting[99] of deepfakes, some bills focusing on other issues considering deepfakes have been introduced.[100]

On state level, the enacted bills mainly target deepfake pornography and electoral/political deepfakes. Protection for candidates for elected office has been provided in both California[101] and Texas,[102] but the latter has a narrower scope only prohibiting visual misrepresentation. Deepfake use in sexually explicit material is targeted in California[103] whereas Virginia prohibits the dissemination of such material if the disseminator has the intent to 'coerce, harass,

[96] David Greene, 'We Don't Need New Laws for Faked Videos, We Already Have Them' (*Electronic Frontier Foundation,* 13 February 2018) <www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them> accessed 27 June 2021.

[97] Unsolicited Explicit Images and Deepfake Pornography Bill waits for second reading in the House of Commons (UK) <https://bills.parliament.uk/bills/2921> accessed 1 August 2021; the EU proposed an Artificial Intelligence Act that would require disclosures of deepfake technology's use: Commission, 'Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative acts' COM (2021) 206 final.

[98] S.1348, 116th Congress (2019-2020) A bill to require the Secretary of Defense to conduct a study on cyberexploitation of members of the Armed Forces and their families, and for other purposes.

[99] H.R.6395, 116th Congress (2019-2020) National Defense Authorization Act for Fiscal Year 2021; H.R.4355, 116th Congress (2019-2020) Identifying Outputs of Generative Adversarial Networks Act; S.2065, 116th Congress (2019-2020) Deepfake Report Act of 2019.

[100] H.R. 3230, 116th Congress (2019-2020) Defending Each and Every Person from False appearances by Keeping Exploitation Subject to Accountability Act of 2019; S. 2559, 117th Congress (2021-2022) Deepfake Task Force Act; H.R.6088, 116th Congress (2019-2020) Deepfakes in Federal Elections Prohibition Act.

[101] AB-730 Elections: deceptive audio or visual media.

[102] S.B.751 Texas Capitol.

[103] AB-602 Depiction of individual using digital or electronic technology: sexually explicit material: cause of action.

or intimidate'.[104] A New York bill also bans deepfake pornography, but the other part of the bill distinguishes it from the other states.[105] The part unique to New York establishes the right to protect one's likeness from unauthorised commercial exploitation for 40 years postmortem, meaning that under New York law deceased individuals cannot be brought back to life (eg with the aid of deepfake technology) without the consent of the decedent's heirs or assignees.
Actions to legislate deepfakes have been taken on the other side of the globe as well. From January 2020 onwards, a regulation introduced by Cyberspace Administration of China requires that any deepfake content is adequately labelled.[106] Additionally, if any deepfake content is not properly labelled, the regulation requires the information service providers to immediately stop the contents' transmission.[107]

The fact that initiative has been taken for legislating deepfakes shows that the urgency to do something for the widespread phenomenon is acknowledged by states. Still, these laws that have passed are somewhat inadequate, and can end up being of no help for the depicted individuals. The defects of these laws stem from their narrow scope – whether the scope is limited by deepfake definition or by the area in which the law applies – and the geographical borders which the laws are limited by. Therefore, I claim that a global initiative is needed, and the scope needs to be broad enough in order to cover areas that are yet to be realised.

2.7 Deepfake Detection

One response to the unwished use of deepfake technology has been to develop different detection methods to assess the authenticity of the content. Moreover, since we do not yet possess consensus regarding the labelling or any other authenticity indicators in creating any audio, image, or video files, the identification needs to be done afterwards by trying to spot the fakes.

The current deepfake detection tools are used to identify whether a file (most often an image or a video) is a deepfake, or they focus on image attribution, ie identifying whether a file was

[104] Virginian Code, Unlawful dissemination or sale of images of another; penalty, s. 18.2-386.2.

[105] S.5959D/A.5605-C New York.

[106] Chinese Provisions on the Management of Online A/V Information Services, Art 11. An unofficial translation available at <www.chinalawtranslate.com/en/provisions-on-the-management-of-online-a-v-information-services/> accessed 8 August 2021.

[107] ibid, Art. 12.

created by any deepfake models.[108] Deepfake detection approaches include,[109] inter alia, reverse engineering and digital fingerprints,[110] attentional visual heartbeat,[111] optical flows,[112] convolutional traces,[113] and exploiting face warping artifacts.[114] There are already companies providing these detection services,[115] and new companies are entering the market to provide authentication tools as well.

Nevertheless, the lack of consensus does not mean that there are no options for ex ante authentication. Whereas detection functions ex-post, authentication tools are used prior the content is distributed, more specifically when the content is created (or, in many cases, captured). Authentication tools and applications (such as controlled capture tools) are already provided for corporations to protect themselves from synthetic media.[116] Blockchain technology has also been suggested for a way of authentication,[117] but before we are more familiar with the environmental burdens blockchain mining causes, I emphasise that we should be cautious when opting for blockchain technology.

[108] Oliver Peckham, 'New AI Model From Facebook, Michigan State Detects & Attributes Deepfakes' (*Datanami*) <www.datanami.com/2021/06/25/new-ai-model-from-facebook-michigan-state-detects-attributes-deepfakes/> accessed 28 June 2021.

[109] For explanation, see Appendix III.

[110] Xi Yin and Tal Hassner, 'Reverse Engineering Generative Models From a Single Deepfake Image' (*Facebook AI,* 16 June 2021) <https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image> accessed 28 June 2021.

[111] Hua Qi and others, 'DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythm', in *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)* (Association for Computing Machinery 2020) 4318.

[112] Roberto Caldelli and others, 'Optical Flow Based CNN for Detection of Unlearnt Deepfake Manipulations' (2021) 146 Pattern Recognition Letters 31.

[113] Luca Guarnera and Sebastiano Battiato, 'Fighting Deepfake by Exposing the Convolutional Traces on Images' (IEEE Access 2020).

[114] Yuezun Li and Siwei Lyu, *Exposing DeepFake Videos By Detecting Face Warping Artifacts* (arXiv 2018) available at <https://arxiv.org/abs/1811.00656v3> accessed 28 June 2021.

[115] Sentinel provides AI-based detection services, Sentinel <https://thesentinel.ai/>.

[116] For example: Truepic <https://truepic.com/technology/>.

[117] Haya Hasan and Khaled Salah, 'Combating Deepfake Videos using Blockchain and Smart Contracts' (2019) 7 IEEE Access 41596.

THREE – A Legal Headache

Since the target is not to create an absolute ban for the use of deepfake technology nor deepfakes, the regulative framework needs to be carefully drafted for it to be feasible. A plethora of obstacles needs to be considered for the framework to achieve broad acceptance amongst different stakeholders. Furthermore, deepfakes cannot be targeted merely on a national level due to the borderless nature of information distribution online. Therefore, the framework needs to acknowledge different jurisdictional backgrounds for the proposal to be normatively appealing and constitutionally permissible.

3.1 Definitions

In legal language, definitions always have certain adversities they need to overcome. The difficulty of capturing the essence of a concept is not common only in modern technologies, but a struggle in the legal field throughout history.[118] These linguistic difficulties stem from the fact that too narrow a definition will leave certain actions or technologies outside the legislations' scope, whereas having too broad of a wording might make the rule too vague and thus downplay its applicability. Decision to not define deepfakes at all guarantees uncertainty by virtue of diverse legal professionals' different interpretations. If the interpretations vary, it could shift the disputes' focus more on defining whether something is a deepfake rather than addressing the disputed problems and harms created by deepfakes.

The chosen vocabulary plays an important role in the definition. Especially if words such as 'harm' or 'intention' (eg deepfakes that create *harm* or that the creator or distributor *intended* to cause harm) are included, the burden of proof is most likely shifted to the plaintiff. The plaintiff would need to prove that there was actual or foreseeable harm caused by the deepfake, or that the creator/distributor intended the content to cause harm. Moreover, if an element of intent is included in the definition, less deepfakes might fall within the definition's scope, if these were created merely for personal entertainment.[119]

Whereas deepfakes are ever so easy to make, an additional issue enters the discussion. This issue relates to whether the ones making deepfakes duly understand that they are, indeed,

[118] For example, in the UK copyright law, a proper definition for 'in public' is still missing. Makeen Makeen, 'Rationalising performance "in public" under UK copyright law' (2016) 2 I.P.Q. 117.
[119] See Ayuub (n 84).

creating deepfakes.[120] Consider, for example, applications that have the capability to be used for creating deepfake content;[121] a question arises whether every app-user understands that they are creating content that is false or falsified.[122] This discussion could be avoided by including a threshold for intent, but as argued above, proving the intent would create an additional burden. Thus, I suggest that instead of intent, a clear definition of what is allowed and prohibited is used.

Moreover, we need to pay attention to the fact whether we separate deepfakes (as the output) from the technology used to produce them, or do we group both the product and the technology together under the same concept. In my opinion, it is important to separate these two; the technology for generating deepfakes in itself is not causing nor producing any positive or negative outcomes, rather it is the human actor who employs the technology and makes the choice of what type of outcome is generated. Since deepfakes have several dimensions, the different steps contributing to these varied stages need to be adequately realised in order to determine who to hold accountable. It follows that there needs to be definitions for the creator (and whether that is the application/program creator or the creator of the deepfake), distributor, and provider.

3.2 Legal Ethics

No technology exists in a void. Technology, as well as deepfakes and the technologies used for creating them, exist within a social context. Inherently, technology is value-neutral, and as put by sociologist Adam Alter, 'tech isn't morally good or bad until it's wielded by the companies that fashion it for mass consumption'.[123] It is the social context and the way individuals choose to employ deepfakes that play the leading role in the conversation regarding the ethical aspects of deepfakes.

[120] For general discussion regarding knowledge, see, for example, Jessica Ice, 'Defamatory Political Deepfakes and the First Amendment' (2019) 70(2) Case W.Res.L.Rev. 417, 449.

[121] For example, Snapchat has face-swap filters and using that might not culminate to actual knowledge of creating fake content.

[122] Keeping in mind the legal principle of *Ignorantia legis neminem excusat*.

[123] Adam Alter, *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked* (Penguin Books 2018) 8.

Legal ethics regarding deepfakes can be roughly divided into two categories. The other focuses on responsible AI practices as a broader topic,[124] and the other pays attention to the question of what is ethical to create by using deepfakes. In relation to deepfakes, responsible AI practices highlight two main elements; the importance of including ethical consideration in the process of developing (deepfake) technologies that generate media, and developing new methods that enable the detection of deepfakes.[125] Different forums that teach and provide courses on deep learning strive to embed ethics as a part of their programmes.[126] If it becomes the norm to include ethical discussion and strive for responsible AI practices, a greater focus can be placed on the discussion of the ethical use of deepfakes.

On one hand, deepfakes can be linked to deception.[127] Because deception compromises norms of truthfulness, it is therefore considered harmful for trust and social relations,[128] hence, deepfakes can have morally and ethically problematic attributes. The discussion on the ethical implications of deepfake technology has touched upon the use of deepfakes for sabotage and blackmail,[129] ideological manipulation[130] and elections,[131] and incitement to violence.[132]

[124] See, for example, 'Responsible AI Practices' (*Google AI*) <https://ai.google/responsibilities/responsible-ai-practices/> accessed 19 August 2021; 'Operationalizing responsible AI' (*Microsoft.com*) <www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5> accessed 19 August 2021.

[125] Bruno Sch_, 'Ethical Deepfakes' (*Bloom AI,* 22 December 2020) <https://medium.com/bloom-ai-blog/ethical-deepfakes-79e2e9eafad> accessed 24 July 2021.

[126] For example FastAI: Rachel Thomas, 'AI Ethics Resources' (*Fast.ai,* 24 September 2018) <www.fast.ai/2018/09/24/ai-ethics-resources/> accessed 24 July 2021; MIT, 'MIT Media Lab' <www.media.mit.edu/users/login/?next=%2Fcourses%2Fthe-ethics-and-governance-of-artificial-intelligence%2F> accessed 23 July 2021.

[127] For example, Dr Thomas King deems Google's AI/deepfake audio assistant as deceiving. 'As for whether the technology itself is deceptive, I can't really say what their intention is — but… even if they don't intend it to deceive you can say they've been negligent in not making sure it doesn't deceive…' Natasha Lomas, 'Duplex Shows Google Failing at Ethical and Creative AI Design' (*TechCrunch*, 10 May 2018) <https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/> accessed 19 July 2021.

[128] Adrienne de Ruiter, 'The Distinct Wrong of Deepfakes' (2021) Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-021-00459-2> accessed 28 August 2021.

[129] Robert Chesney and Danielle Citron, 'Deepfakes and the New Information War' (2019) 98(1) Foreign Affairs 147, 147–155.

[130] John Fletcher, 'Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-fact Performance' (2018) 70 Theatre Journal 455.

[131] Nicholas Diakopoulos and Deborah Johnson, 'Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections' (2020) 23(7) New Media & Society 2072.

[132] Citron and Chesney (n 10) 1753–1819.

Recently the ethical dilemmas of using deepfake-generated content as a part of film production have surfaced when Anthony Bourdain's voice was synthetically generated to read three lines in the film *Roadrunner: A Film About Anthony Bourdain*.[133]

Ethics have to be considered when asking whether bringing deceased individuals back to life without their (or their heiress'/assignees') consent is acceptable, and how using their likeness could possibly impact the reputation of these people.[134] Even if in the entertainment industry the use of deepfakes is not inherently bad, the moral and ethical evaluation of the specific use depends on whether the depicted individual(s) would consent to the way in which they are represented. Or, as Adrienne de Ruiter put it, the ethical rightfulness can be assessed on 'whether the deepfake deceives the viewers, and on the intent with which [it] was created'.[135] Therefore, the ethical aspects should be borne in mind while drafting the framework, as the legislation should accommodate the society's views of accepted use of deepfakes.

3.3 Cross-border nature of the Internet

The internet does not know geographical borders; we can easily scroll through content in the UK published by another user in Thailand. Furthermore, by using VPNs[136] and other encryption methods, users can change their digital location, thus making the user's physical location even less relevant. Because of the possibility of placing oneself anywhere online, and having access to content across the globe, it is substantial that the legal framework is not limited by state borders. If there are only few nations that take upon the initiative to regulate deepfakes through domestic laws, it will be a far cry from an efficient and well-functioning legal tool for addressing the deepfake problems. Moreover, the states with domestic regulations could not enforce those laws or take action against actors residing outside the state, and even hosting the content abroad could prevent triggering the domestic clauses targeting deepfakes.

[133] Adrian Horton, 'The Bourdain AI Furore Shouldn't Overshadow an Effective, Complicated Film' *The Guardian* (21 July 2021) <www.theguardian.com/film/2021/jul/21/anthony-bourdain-documentary-roadrunner-ai-deepfake-furore> accessed 22 July 2021; Helen Rosner, 'The Ethics of a Deepfake Anthony Bourdain Voice' *The New Yorker* (17 July 2021) <www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice> accessed 22 July 2021.

[134] 'The Ethics of Deepfake Aren't Always Black and White' (*The Next Web,* 16 June 2019) <https://thenextweb.com/news/the-ethics-of-deepfakes-arent-always-black-and-white> accessed 22 July 2021.

[135] de Ruiter (n 128).

[136] Virtual Private Networks.

3.4 Multiple stakeholders

Deepfakes are not a threat that have a single clearly defined target or use. Instead, in the negative sphere, deepfakes target the society as a whole, its democratic and political arenas, individuals, corporations, and national security. To be able to protect societies and their citizens, multiple stakeholders' engagement is a prerequisite. The bare minimum of stakeholders that I identify as essential are platform providers (eg Facebook,[137] Google,[138] TikTok, Twitter), software developers (especially those developing on Android and iOS), state actors and regional groups (EU, Europol), and industry associations (eg MPAA,[139] SAG-AFTRA[140]). For the framework to be normatively appealing for the representants of the different fields, input is required from all these stakeholders.

Including multiple stakeholders from the very beginning of the process could decrease the dissatisfaction and amount of lobbying[141] at the time of publishing the initial proposal, when already a considerable number of hours and money have been used.

Depending on the clauses the framework contains, different input from the stakeholders is needed to understand what would be accepted and what could be feasible to develop. For the industries that employ deepfakes for benign use, the industries should be consulted during the drafting process. Understanding the possibilities and the ways in which the entertainment industry[142] aims to use deepfake technology is crucial in the drafting process, so that the framework does not amputate the legs of a promising way of development for deepfake technology.

If labelling was a requirement for all online platforms to implement, input from platform providers and their developers could provide insight on labelling standards and methods. The drafters need to understand what is wished from the branches and what can be done, since the algorithms and scripts should be sophisticated enough to recognise when the content is truly a

[137] Including Instagram and WhatsApp.

[138] Including YouTube.

[139] Motion Pictures Association America.

[140] Screen Actors Guild - American Federation of Television and Radio Artists.

[141] For example, MPAA and Disney advocated against the New York bill. Katyanna Quach, 'New York State is Trying to Ban 'Deepfakes' And Hollywood Isn't Happy' *The Register* (12 June 2018) <www.theregister.com/2018/06/12/new_york_state_is_trying_to_ban_deepfakes_and_hollywood_isnt_happy/> accessed 15 July 2021.

[142] And other industries identified in 1.1.

deepfake, and not to label content that has merely slight edits.[143] Moreover, if the framework imposes responsibilities on social media platforms, these platforms are thereby encouraged to draft their terms of services and general user policies using the same universal standards. I suggest these universal standards so that certain online behaviour in relation to deepfake content and its creation becomes a norm regardless of the platform.

Another thing to consider is the possibility to require an authentication of all online content. For content authentication to be obtainable for any user, understanding the technicalities of software and application development for devices capable of recording audio or visual and audiovisual material is important in the drafting process.

3.4.1 Platform responsibility

Platform responsibility is highlighted as a separate section for few reasons. First, online platforms are owned by private corporations, and the First Amendment is not applicable to those. Hence, these corporations have more freedom to decide what can be distributed on their platforms.[144] Second, Section 230 of the US Communications Decency Act provides platforms with immunity of third-party content liability. Under the Good Samaritan protection provided in said Section, the platforms can exercise content removal policies and labelling while still being protected from civil liabilities in this regard. Third, corporations that own these platforms are able to shape the published content through their terms of service, by simply choosing to allow only certain types of content or behaviour on their platforms.

Social media platforms that host notable amounts of video, audio, and image content – and thus are in the focus when it comes to deepfakes – include YouTube, Instagram and Facebook, TikTok, Onlyfans, and Twitter. In addition to social media platforms, I consider Google[145] to have possibilities to impact the search results. For example, if one is doing a Google search for deepfake pornography applications or forums, there are plenty of sites that pop up directly on the first page. Even if Google has put money into deepfake detection work,[146] it could also impact the selection of sites that appear when performing a search, by pushing the sites lower

[143] Edits eg improved lightning, trimming of audio clips.

[144] See discussion in 3.5.

[145] Google here refers to the search engine.

[146] Nick Dufour and Andrew Gully, 'Contributing Data to Deepfake Detection Research' (*Google AI blog,* 24 September 2019) <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html> accessed 20 March 2021.

in the search results or tweaking their search engine optimisation algorithms. Lastly, sites that host deepfake pornography should act responsibly and ascertain that the content does not account to revenge porn, unconsented porn, nor any other types of porn classified illegal under national laws.

Even if platforms can take action in this realm, it should not be forgotten that these platforms are owned by private corporations that are incorporated for making a profit and serving, for example, the advertisers. Social media platforms are not public forums nor public services.[147] It is, of course, the flipside of the coin that by relying on (social media) platforms to take responsibility, we provide them with a tighter grip of our lives and rights online.

3.5 Free Speech and Freedom of Expression

Regardless of the implicit falsity of deepfakes, they do implicate the freedom of speech and freedom of expression. And when the fundamental rights of freedom of expression and freedom of speech are limited, the restrictions to them must be proportional and provided for by law.[148] The strength of these fundamental freedoms is the most frequently used justification for not to regulate deepfakes,[149] and scholars have noted the possibilities of deepfake regulations to 'engage in censorship of free expression online'.[150]

The ECtHR has illustrated that the freedom of expression online can be limited,[151] while establishing that the general principles covering offline publications also apply online.[152] Furthermore, the Court has considered the ability for information to spread online,[153] while

[147] Jakob Miller, 'Is Social Media Censorship Legal?' *The Echo* (22 February 2021) <www.theechonews.com/article/2021/02/kdyntyejlxplpeh> accessed 24 July 2021.

[148] Ian Cram, *Contested Words: Legal Restrictions on Freedom of Speech in Liberal Democracies* (Routledge 2016).

[149] Holly Hall, 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27(1) Catholic University Journal of Law and Technology 51.

[150] Sharon Franklin 'This Bill Hader Deepfake Video Is Amazing. It's Also Terrifying for Our Future' *New America* (13 August 2019) <www.newamerica.org/oti/in-the-news/bill-hader-deepfake-video-amazing-its-also-terrifying-our-future/> accessed 17 August 2021.

[151] *Willem v France* App no 10883/05 (ECtHR, 16 July 2009); *Féret v Belgium* App no 15615/07 (ECtHR, 16 July 2009), para 80.

[152] *Aleksey Ovchinnikov v Russia* App no 24061/04 (ECtHR, 16 December 2010) paras 49-50; *Renaud v France* App no 13290/07 (ECtHR, 25 February 2010) para 40; *GRA Stiftung Gegen Rassismun und Antisemitismus v Switzerland* App no 18597/13 (ECtHR 9 January 2018); *Féret v Belgium* (supra n16) para 78; *Willem v France* (supra n16).

[153] *Delfi AS v Estonia* App no 64569/09 (ECtHR, 10 October 2013).

speculating whether offensive communications have less impact online due to the internet's attributes.[154]

The First Amendment to the US Constitution states that the 'congress shall make no law (…) abridging the freedom of speech'. Even if the wording of the First Amendment prohibits government actions to abridge free speech, it does not apply to private companies and their rights to compile their terms of service policies to censor deepfake content. Although, ECHR provides the right to information as an integral part of the freedom of expression. Hence, automatically blocking all deepfake content[155] or banning access from deepfake content creators and/or disseminators could face hardship regarding the framework's feasibility if contested against human rights. Therefore, the scope of the legal framework needs to be carefully drafted so that it would not be declared unconstitutional in the US and elsewhere.

Lastly, a Harvard professor suggests that the US government can regulate deepfakes without the regulation or a ban being unconstitutional is possible if '(1) it is not reasonably obvious or explicitly and prominently disclosed that they are deepfakes, and (2) they would create serious personal embarrassment or reputational harm'.[156] Alternatively, if prohibited deepfakes are to be classified under unprotected speech, they would then fall outside of constitutional protection.[157] Thus, deepfake regulations do not need to be declared automatically unconstitutional.

3.6 Anonymity

Online anonymity is a double-edged sword. On one hand, it enables broader democratic rights[158] and is linked to 'people's willingness to engage in debate on controversial subjects in the public sphere'.[159] Online anonymity also protects individuals from offline violence[160] and

[154] *Magyar Tartalomszolgáltatók Egyesülete and Index.Hu ZRT v Hungary* App no 22947/13 (ECtHR, 5 January 2016).

[155] Blocking of access might limit the right to freedom of expression too broadly. See, for example, *Ahmet Yıldırım v Turkey* App no 3111/10 (ECtHR, 18 December 2012).

[156] Cass Sunstein, 'Falsehood and the First Amendment' (2020) 33(2) Harv.J.L.&Tech. 387, 421.

[157] *United States v. Alvarez* (2012) 567 U.S. 709, at 719, 725.

[158] *McIntyre v Ohio Elections Commission* (1995) 514 U.S. 344.

[159] UNGA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue' UN Doc A/HRC/17/27 (2011), 15.

[160] Robert Bodle, 'The Ethics of Online Anonymity or Zuckerberg vs "Moot"' (2013) 43(1) Computer and Society 22.

protects whistle-blowers.[161] On the other hand, online anonymity enables deepfakes, as well as racist and hateful comments, to flourish due to the lack of accountability and traceability.[162]

If access to online platforms requires ID verification, [163] a billion people would be automatically banned from these platforms, since one billion people do not have an official proof of identity.[164] Even if banning anonymity can have good intentions and ease discovering the deepfakers, it simultaneously creates burdens for access that would either directly or indirectly undermine individuals' rights to access information and thus diminish their ability to participate in public discussion.

I argue that online anonymity is of fundamental importance in securing public and political discourse and assuring whistle-blowers safety. But still, online anonymity greatly decreases the possibilities for deepfakes' targets to find someone to hold accountable for the content depicting them. Hence, interests need to be carefully balanced when addressing the use of anonymity and its impact on deepfakes.

3.7 Intellectual Property Rights

Reflecting on the discussion presented in 2.1, intellectual property rights can also present problems for deepfake regulation, especially copyrights. WIPO has also highlighted problems of deepfake content in the 2019 Draft Issues Paper On Intellectual Property Policy And Artificial Intelligence.[165] WIPO explicitly addresses two problems, the first concerning the copyright of a deepfake, and the other addressing the question whether a 'system of equitable remuneration for persons whose likenesses and "performances" are used in a deep fake' should

[161] Kathleen Clark and Nancy Moore, 'Buying Voice: Financial Rewards for Whistleblowing Lawyers' (2015) 56(5) B.C.L.Rev. 1697; Tanya Marcum, Jacob Young and Ethan Kirner 'Blowing the Whistle in the Digital Age: Are You Really Anonymous? The Perils and Pitfalls of Anonymity in Whistleblowing Law' (2020) 17(1) DePaul Bus.L.J. 1.

[162] Hussein Kesvani 'Abolishing Online Anonymity Won't Tackle the Underlying Problems of Racist Abuse' *The Guardian* (15 July 2021) <www.theguardian.com/commentisfree/2021/jul/15/abolishing-online-anonymity-racist-abuse-id-verification> accessed 16 July 2021.

[163] BCS suggested after the English football team receiving racist comments. BCS, 'Social media must verify users' ID to end online abuse - IT industry poll' (*BCS.org,* 29 April 2021) <www.bcs.org/more/about-us/press-office/press-releases/social-media-must-verify-users-id-to-end-online-abuse-it-industry-poll/> accessed 16 August 2021.

[164] The World Bank, 'ID4D Data: Global Identification Challenge by the Numbers' (*The World Bank Data Catalog,* 2018) <https://datacatalog.worldbank.org/dataset/identification-development-global-dataset> accessed 20 July 2021.

[165] WIPO (n 70).

exist or not.[166] More than providing arguments in the realm to whom copyrights should be allocated to, the main concern in WIPO's paper is whether copyrights should be rewarded for deepfake content at all. Due to the many positive possibilities deepfake technology provides, I argue that there should be a reasonable system for providing intellectual property rights for the creators of deepfakes. It cannot be ignored that such an argument might open the Pandora's box of IP rights attached to AI-generated works,[167] but deepfake related discussion should consider that many deepfakes include creative choices, and thus are not a mere production of AI.

WIPO also presents the idea that the copyrights (if rewarded) should be granted to the inventor of deepfakes, but I do not consider this as a satisfactory solution – this would be the equivalent of Picasso and Braque owning the copyright on all cubist paintings.[168] To provide appropriate protection of the intellectual property rights for those employing deepfake technology in their line of work, multiple stakeholders should be heard in order to determine what should be included in the legal framework. For a regulative approach to be generally accepted across different sectors, associations such as SAG-AFTRA and MPAA should be consulted to avoid any well-intended laws unnecessarily hindering the positive developments of deepfakes and related technologies.

3.8 Enforcement

The issues regarding enforcement are closely related to the cross-border nature of the internet. The key questions to answer include the enforceability of remedies, what would be the suitable institution or enforcing body, and speediness of the processing.

[166] ibid 6.

[167] See, for example, Reto Hilty, Jörg Hoffmann and Stefan Scheuerer, 'Intellectual Property Justification for Artificial Intelligence' in Jyh-An Lee, Reto Hilty and Kung-Chung Liu (eds) *Artificial Intelligence & Intellectual Property* (OUP 2020) 20; Russ Pearlman, 'Recognizing Artificial Intelligence (AI) as Authors and Inventors under U.S. Intellectual Property Law' (2018) 24(2) Rich.J.L.&Tech. 1; Francis Gurry, 'Artificial intelligence and intellectual property: An interview with Francis Gurry' (2018) 114 Intellectual Property Forum: Journal of the Intellectual and Industrial Property Society of Australia and New Zealand 66; Sonali Kokane, 'The Intellectual Property Rights of Artificial Intelligence-based Inventions' (2021) 65(2) Journal of Scientific Research 116; Committee on Legal Affairs, 'Report on intellectual property rights for the development of artificial intelligence technologies' (2020/2015INI) (2 October 2020) <www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html> accessed 24 July 2021; Suebsiri Tawepoon and Pimpisa Ardbroriak, 'Challenges of Future Intellectual Property Issues For Artificial Intelligence' (*Tilleke & Gibbins,* 6 December 2018) <www.tilleke.com/insights/challenges-future-intellectual-property-issues-artificial-intelligence/> accessed 24 July 2021.

[168] Picasso and Braque are considered the inventors of cubism. Tate, 'Art term – Cubism' (*Tate.org)* <www.tate.org.uk/art/art-terms/c/cubism> accessed24 July 2021.

Depending on the classification of the legal framework – does it fall under civil or criminal law provisions –, the institution needs to be such that possesses the authority to operate in the chosen area of law. One suitable option could be a globally operating organ (an administrative one, possibly created under a UN mandate due to the universal presence and acceptance of the intergovernmental organisation) that would have regional offices in different geographical areas[169] overseeing the national application and enforcement of the legal framework. The benefit of such approach would be allocating more resources on processing the cases, thus making the processing more efficient as opposed to a single institution handling all the procedures.

If overseeing the implementation as well as the enforcement of the legal framework were concentrated under one institution, the rules and procedures would be close to universal. This, in turn, would create legal stability and predictability, and the processes for seeking remedies and removals of deepfake content would not be stopped merely for the lack of jurisdiction across geographical borders.

3.8.1 Retroactivity

The general principle of non-retroactivity is commonly agreed upon and included in the constitutions across different legal traditions.[170] Whether the proposed framework is categorised under criminal or civil law has an effect on how retroactivity is addressed. Most often the domestic constitutions and criminal laws prohibit retroactivity regarding crimes, and the laws that states have adopted with retroactive effect are usually outside of the criminal sphere.[171]

Addressing retroactivity within the legal framework concerning deepfakes is problematic in two main areas. First, it needs to be adequately defined whether the framework targets the creation, publication, or distribution of the content. If the content was generated prior the framework enters into force, but published only after the framework's applicability, the framework needs to indicate whether the depicted individuals would have any plausible actions to take. Second, it might take time until the depicted individuals become aware that their

[169] Exemplary division: Europe/EU, Americas, Middle East, Africa, Asia, and Oceania.

[170] See, for example, US Constitution Art. I § 9; Gambian Constitution of the Republic of Gambia Art. 24(5); Swedish Constitution, Regeringsfomen Art. 10; Russian Constitution of the Russian Federation Arts 54, 57.

[171] Israeli Nazis and Nazi Collaborators (Punishment) Law 5710–1950 (1949-1950) 4 LSI 154; UK War Crimes Act 1991; Australian Taxation (Unpaid Company Tax) Assessment Act 1982.

likeness or images have been used. If the framework did not apply to content retroactively, these individuals would be left with no aid from the statutory provisions. Clauses that provide tools for the individuals to get the content removed could be useful in regulating deepfake content even if the deepfakers and disseminators could not be held accountable due to the constitutional barriers of retroactivity.

3.9 Exceptions

Where there are rules, there are exceptions. Copyright laws include fair use, violent crimes recognise self-defence, and privity of contract doctrine acknowledges exceptions that allow third parties to have rights and/or obligations.[172] That is to say, the legal framework for deepfake content needs to provide room for certain situations or types of use when otherwise prohibited action would be allowed.

Exceptions could include parody and satire, as well as public interest, provided that it would be sufficiently defined to avoid becoming too broad a loophole. Public interest exceptions could for example allow otherwise malicious or harmful dis- or misinformation, if such content informed the public of harms that deepfakes can create, or contributes to research, or broader political or public discussion.

[172] See, for example, *Beswick v Beswick* [1967] UKHL 2, [1968] AC 58.

FOUR – The Proposal

We are willing to accept that we cannot extend legal rules of steam locomotives[173] to cover self-driving cars,[174] so we should not insist on tackling deepfake problems with the aid of existing laws that cover image rights and defamation. Hence, I suggest a legal framework specifically tailored for deepfakes, together with non-legal recommendations for a healthier online environment.

4.1 The Statutory Toolbox

Throughout this opinion it has been highlighted how technology does not exist in a void. Likewise, deepfake content does not have a clearly defined target audience, instead, deepfakes continuously increase their presence in new areas. Therefore, the legal framework needs to be globally acknowledged and enforceable, commonly agreed upon by all the relevant stakeholders, and sufficiently flexible so that new ways of developing deepfake content are still captured by the framework.

4.1.1 Preamble

The legal framework should include a preamble. The drafters included in the process should **outline the desired outcome** of the legal tool, and thus provide the guidelines for those interpreting the framework. In this case, the aim is **not to create an absolute ban** for the use of deepfake technology nor deepfakes, but instead to **limit the malicious and harmful deepfake content and increase transparency**.

4.1.2 Definitions

Formulating all the necessary definitions is a difficult but necessary part of creating the framework. I argue that the framework should **not define the deepfake technology itself**, since that can make the framework run behind technical development. Moreover, certain components

[173] UK, Locomotives Act 1861.

[174] See, for example, UK Law Commission's project on legal framework for automated vehicles UK Law Commission, 'Automated Vehicles' (*Law Commission*) <www.lawcom.gov.uk/project/automated-vehicles/#related> accessed 20 March 2021.

of deepfake technology are included in several image-processing tools,[175] and the inclusion of certain technological attributes under deepfake-definition can make the framework function against its aim and lead to unwished outcomes. Still, certain **attributes of the content need to be defined** so that it can be identified as a deepfake. **The definition for deepfakes** could be drafted among the lines of it being **synthetic media** (auditory, visual, or audiovisual) generated by using deep or machine learning technology and appearing to be authentic to an ordinary person.

Unauthorised deepfakes refer to such that use existing individuals'[176] facial, vocal, or bodily attributes to impersonate them without authorisation. **Malicious deepfakes** are described as having the ability to deceive audiences and/or to influence perceptions of reality, or depicting individuals in scenarios that they would not consent to be presented in.

Not only does the framework need to define the deepfakes themselves, but the **actors involved** as well. These are the creator, distributor, host, and the depicted individual. Since the aim is not to target the technology, there should be distinction between the **creator of the technology** or applications, and the **creator of the content** itself This could be achieved by referring to the content creator as a **'deepfaker'**.[177] Furthermore, **disseminator** covers those individuals or accounts[178] that make the content available to the public, distribute it, or host the content on their platforms.

The depicted individual refers to either a private or public person whose likeness or recognisable attributes are used in a deepfake. It should also be noted that the depicted individual does not refer to a person whose data has been used to train a dataset.[179]

Outside of legal language, it is possible to refer to victims of deepfakes, but within the framework the term 'victim' is not used, as it could be attributed to criminalised actions and criminal law.

[175] 'Deep machine learning is used in so many image based tools, upscaling, content-aware resizing, recoloring, masks for compositing. Not only would it be an impossible task to determine which tools were used in the creation chain, outlawing media synthesized using algorithms would ban Welcome to Chechnya, and any art, ad, video, image which used a machine learned tool.' Email from Ryan Laney* to the author (12 July 2021).
*Ryan Laney is the founder of Teus Media (www.teus.media) and an expert in visual effects. Laney created the deepfake visual veil for the documentary film *Welcome to Chechnya,* prior which he has worked with visual effects for films such as Green Lantern, Hancock, and Spider Man 3.

[176] Meaning fully artificial depictions, such as those generated by eg thispersondoesnotexist.com/, do not count as unauthorised since there is no impersonating or depiction of existing individuals' likeness.

[177] This term is to be considered as a noun, and not as the comparative form of an adjective.

[178] Accounts refer to the possibility to make legal persons accountable.

[179] Regarding datasets, see above footnote 3.

4.1.3 Prohibited and allowed output

Deepfake **technology** itself, its development, or providing the technology's source code (or modifications to it) to the public shall **not be prohibited**. Merely **using the technology** through automation does not account for a prohibited action, since it lacks decisive input. We do not want to hold machines or computers accountable, but the persons who utilise deepfakes in a way that bears negative attributes should be held responsible. **Developing applications** that can be used for creating prohibited output shall be allowed, because banning application development could hinder technological advances.

Creating deepfakes shall be prohibited, if the creation includes any choices made by a person or persons,[180] can be reasonably anticipated or deemed malicious,[181] and is not authorised. Creating is defined as any process that includes choices to bring a form of media that is not authentic into existence, regardless of whether the creator uses pre-made applications, builds the software, or utilises any other means.

Dissemination of deepfakes includes publication, distribution, and hosting. **Publication** of unauthorised deepfakes or deepfakes whose creation is restricted shall be prohibited. In this context, publishing or publication means making the content available to the public, and includes any uploads to online platforms that are deemed to be generally accessible to the public. Publication of fully automated deepfakes can be prohibited if the publication includes a narrative that has, or could have, malicious attributes. **Prohibited distribution** covers any sharing of prohibited content for reasons of making it reach additional audiences. The distribution shall also be prohibited in situations when adequate disclosures of the content's source have been removed or altered.[182] The prohibition to distribute deepfakes shall not deteriorate the possibilities to exercise one's right to express their ideas or opinions, or to hinder any research or other type of fair use.

[180] Choices can be eg, the choice of: training data, impersonated person(s), captured scenario. It does not matter whether choices are made by legal or natural persons.

[181] Malicious in the sense as provided above in 4.1.2.

[182] Altered in this content means, eg, when a watermark has been re-sized smaller or tried to be diluted so that it would be harder to recognise the content's origins.

Allowed deepfake creations shall, inter alia, include any deepfakes that emerge without person(s) making any choices,[183] are authorised by the individual whose likeness (or any other recognisable attributes) are used, are authorised and used in the entertainment, medical, or artistic fields, and anticipated not to bear malicious attributes.

The line between the prohibited and allowed output is not determined by 'intent' or 'harm'. If these terms were used, it could create additional burdens to the person taking action against the deepfake(s), by increasing the threshold that would need to be proved beyond reasonable doubt.

4.1.4 Scope

Geographically, the framework should be **applicable and enforceable everywhere**. This would require sufficient peer-pressure for all states to become signatories of the framework. If the coverage is not broad enough, the framework will not attain its desired outcome due to the lack of enforcement.

The framework should capture actions by both **natural and legal persons**. If legal persons were excluded, it would provide opportunities for some actors to hide behind the corporate veil.

Actions captured within the framework shall be determined in a manner that the framework will not hinder the welcomed developments of deepfakes and deep learning technologies. The framework shall live up to its aim, and indicate that deepfakes themselves are not inherently problematic nor morally or ethically bad. It should be built into the framework that there is no willingness to stop deepfakes, but to diminish the negative outcomes individuals and society can experience.

4.1.5 Stakeholders and required actions

To secure the constitutionality and feasibility of the framework, the framework shall consider the relevant stakeholders and how to guide the general use of deepfake technology to the direction that advances the framework's aim.

[183] For example, Thispersondoesnotexist <thispersondoesnotexist.com>.

4.1.6 Attributable rights

To express the understanding of the benign possibilities deepfake technology can be used for, the framework shall **express support for the possibility to attribute rights for deepfake creators**.

It should be spelled out that there is a possibility to attribute rights in relation to deepfakes within **intellectual property laws**, as well as the possibility to **contractually agree upon the use, release, and/or licensing** of one's likeness, identity, or recognisable attributes to be used in deepfakes.

4.1.7 Remedies

The framework shall **establish a way to seek remedies** for the persons whose recognisable facial, vocal, or bodily attributes have been used without their authorisation, whether for malicious or non-malicious use. **Remedies shall include monetary awards, injunctions, and content removals**. For malicious unauthorised deepfakes, awards shall be higher that benign unauthorised deepfakes.

When imposing monetary awards, the type of the content (whether eg deepfake pornography, political disinformation, evidence tampering) shall be taken into account when deciding whether compensation for emotional distress or financial loss should be awarded.

Injunctions will most likely ask the deepfakers and disseminators to stop further dissemination of the content. Securing the possibility for content removal will allow the person(s) to get the content removed by the platform provider if the disseminator will not do so.

4.1.8 Exceptions

Exceptions shall be granted. Exceptions shall be aligned so that the otherwise prohibited deepfakes are allowed if those **enhance public and democratic discourse,**[184] **appreciate creativity,**[185] **or serve the public interest**.

4.1.9 Establishing the organ

The legal framework shall establish an organ and provide a mandate for the organ to operate.

[184] Eg fair use.

[185] Eg satire and parody.

There shall be **one globally operating organ** that has regional offices in different geographical areas overseeing the national application and enforcement of the set of legal rules. There shall be a possibility to appeal any decisions delivered by the organ.

4.1.10 Enforcement

Enforcement shall be done **on a regional level** with the aid of national offices. An adequate appeals procedure shall be guaranteed. The timeframe of bringing action needs to be clarified, and the possibilities of initiating action retroactively shall be covered.

4.2 Non-legal approaches

Deepfakes and the threats they pose are not something that the governments and states can fix themselves. Therefore, I argue that in addition to the legal framework, non-legal tools need to be utilised in order to create an online society that is able to function in the deepfake jungle.

4.2.1 Awareness

Awareness of the existence of deepfakes is crucial, and this is an action that the online users can take themselves. Informative material on deepfakes has started to emerge, and a lot of the material is available for free and without subscriptions.[186] Besides actions taken by individuals, it is important that corporations and state organs increase their awareness on the topic, especially in relation to the security threats that deepfakes pose. By increasing awareness, these big actors can implement and adjust their security procedures to be more suitable for protecting themselves from deepfake-related threats.[187] Increasing awareness can also be encouraged by educational institutions. One example is from Finland, where courses have been launched to increase citizens' media literacy skills to spot fake news.[188]

Another side of the awareness and learning function is to educate what type of content can be used for creating unwished deepfakes, and how this material is available. This invites social

[186] For example, United Arab Emirates published a deepfake guidance report. National Program for Artificial Intelligence, 'Deepfake Guide July 2021' (www.ai.gov.ae 2021) <https://ai.gov.ae/wp-content/uploads/2021/07/AI-DeepFake-Guide-EN-2021.pdf> accessed 16 August 2021.

[187] For example, employee trainings: Alexander Jones, 'Assessing the Real Threat Posed by Deepfake Technology' (*International Banker,* 24 February 2021) <https://internationalbanker.com/technology/assessing-the-real-threat-posed-by-deepfake-technology/> accessed 20 July 2021.

[188] Eliza Mackintosh, 'Finland is winning the war on fake news. What it's learned may be crucial to Western democracy' *CNN* (May 2019) <https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/> accessed 8 August 2021.

media (and other online) platforms to disclose information regarding protection and privacy of user accounts, and how allowing only known people to access one's content can already limit the risk of being a so-called victim of a deepfake attack. Overall, we should not end up in a situation where automatic denial of content's truthfulness is the norm. If such routine becomes dominant, the ill-intended actors would benefit from it, since they could rely on audience's habitual dismissal and act more carelessly.

4.2.2 Marking

To be able to verify, authenticate, and detect deepfakes on somewhat similar terms globally, there needs to be common standards so that the same tools and methods can be utilised regardless of location. Options that exist include blockchain, hashing, watermarking and labelling, metadata and compression information, and faketagging.[189] One rather simple method is SIFT methodology, which encourages the audience to 'Stop, Investigate the source, Find trusted coverage, and Trace the original context'.[190] This methodology emphasises that the more advanced image forensics might make the audience dive into a deeper rabbit hole instead of helping out in assessing the reliability of the content. Other suggestions have shifted the pressure away from the audience and put it on the applications' creators, by proposing safety frameworks for developers.[191] Further discussion on the detection and authentication solutions has addressed the question of access and exclusions.[192] Hence, prior to reaching some level of consensus, the current methods of detection and authentication can be of help in individual situations, but I argue that they will not be the cure for the problems persons targeted in deepfakes are experiencing.

Furthermore, if any detection, labelling, or authentication tools were included as a requirement, excessive negotiations would be needed so that all states could come into an agreement with one (or few) provider(s) for the signatories to use. The predicaments of such decisions are directly linked to the states' economic interests, as each state would undeniably want their country's corporations/actors to get the deal and thus generate revenue for the state.

[189] See Appendix IV.

[190] Sam Gregory, 'Authoritarian Regimes Could Exploit Cries of 'Deepfake'' *Wired* (14 February 2021) <www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/> accessed 2 July 2021.

[191] Henry Ajder and Nina Schick, 'Deepfake Apps Are Here and We Can't Let Them Run Amok' *Wired* (30 March 2021) <www.wired.co.uk/article/deepfakes-security> accessed 2 July 2021.

[192] WITNESS, 'Deepfakes: Prepare Now (Perspectives from Brazil)' (*WITNESS Media Lab*) <https://lab.witness.org/brazil-deepfakes-prepare-now/> accessed 2 July 2021.

Conclusion

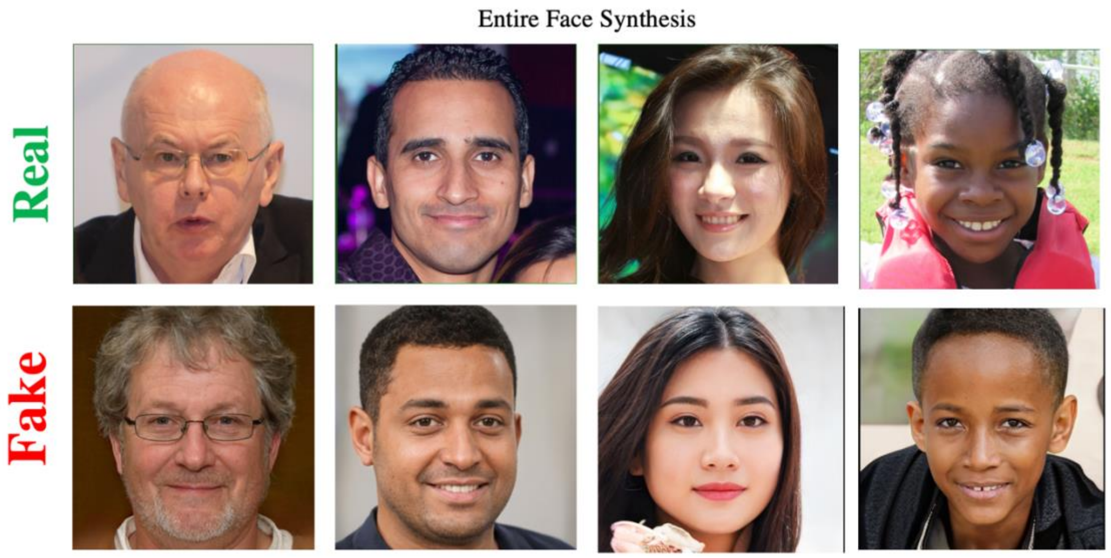
Since November 2017, deepfakes have notably evolved and become more accessible. Deepfakes have already shaped online practices and dissemination of information. Deepfakes provide unlimited opportunities for the well-intended users to exploit new technological developments. At the same time, deepfakes and their presence have provided a considerable expansion of misusing individuals' likeness for malicious purposes, utilising deepfake technologies in a way that cannot be beaten by isolated actions nor current legal rules. As a result, initiatives to create innovative approaches for deepfake regulation, detection, and authentication to combat the malicious users of deepfake technology have become an essential matter for governments and stakeholders globally.

In this opinion, I have proposed to establish a legal framework to battle the ever-growing problem concerning deepfake regulation. A new legal framework that is acknowledged and enforced globally is deemed as a promising instrument in the pursuit for limiting the presence of harmful deepfake content in the online environment. In a broader sense, the legal framework can also set boundaries for allowed use, and steer the behaviour of online users, as well as advance transparency and reliability of digital content. Furthermore, the legal framework is expected to be welcomed due to the many attempts trying to address the deepfake problem.

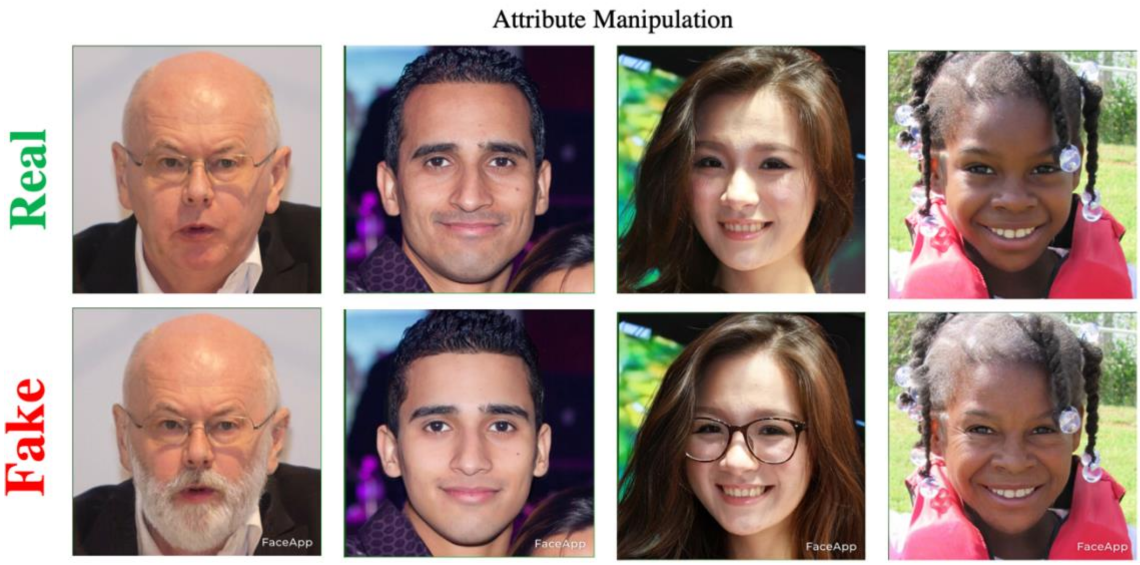
To recapitulate, this legal opinion is written to highlight both the benign and malicious uses of deepfakes and the technologies used for creating them, and to illustrate how existing laws fall short in the battle against unwished forms of deepfakes. Beyond enhancing awareness of the many possible uses of deepfakes and the drawbacks of the current laws, the opinion invites all relevant stakeholders to take an active role in the development of a new legal framework to tackle the modern issue. The legal framework is supported by non-legal approaches, which will provide a safer and more standardised online environment in which to manoeuvre.

Appendices

Appendix I – Examples of different visual deepfakes[193]



Entire face synthesis depicts how real-like images of non-existing persons can be created by using deepfake technology.[194] These fakes have been generated by using GANs.

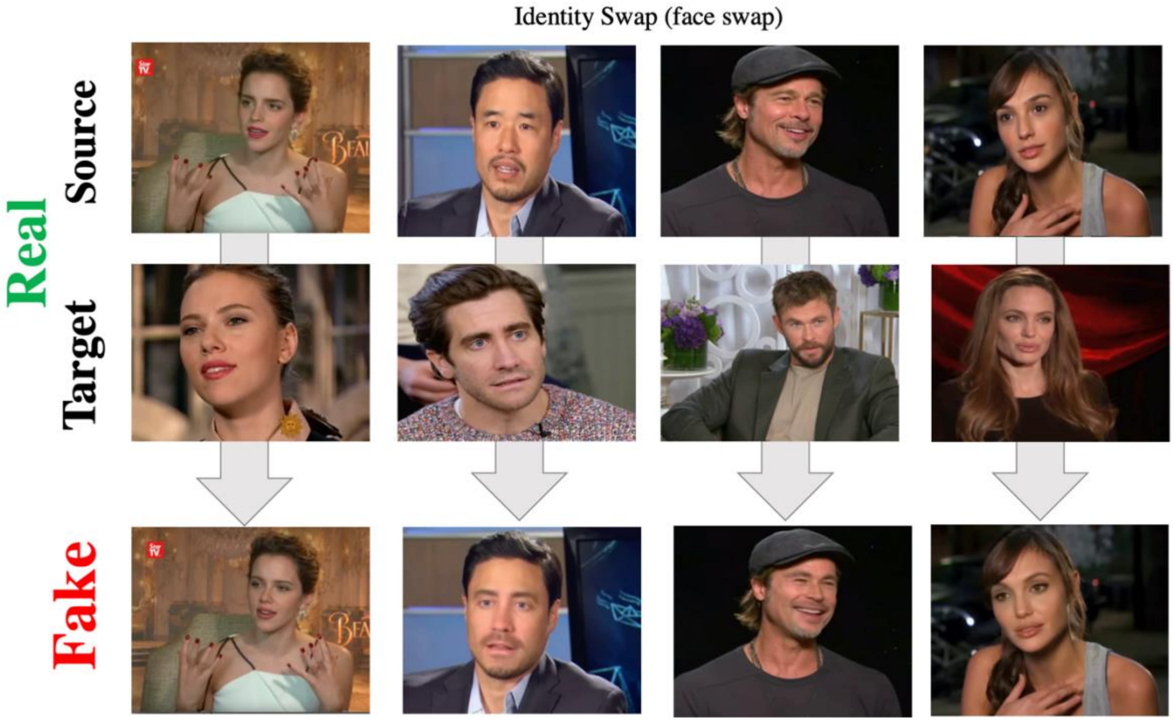


Attribute Manipulation shows how AI generated applications (in this example FaceApp[195]) was used to modify real images of existing individuals. Examples show how facial hair can be added, how aging and de-aging is possible, as well as changing hair style, adding glasses, and accentuating smile.

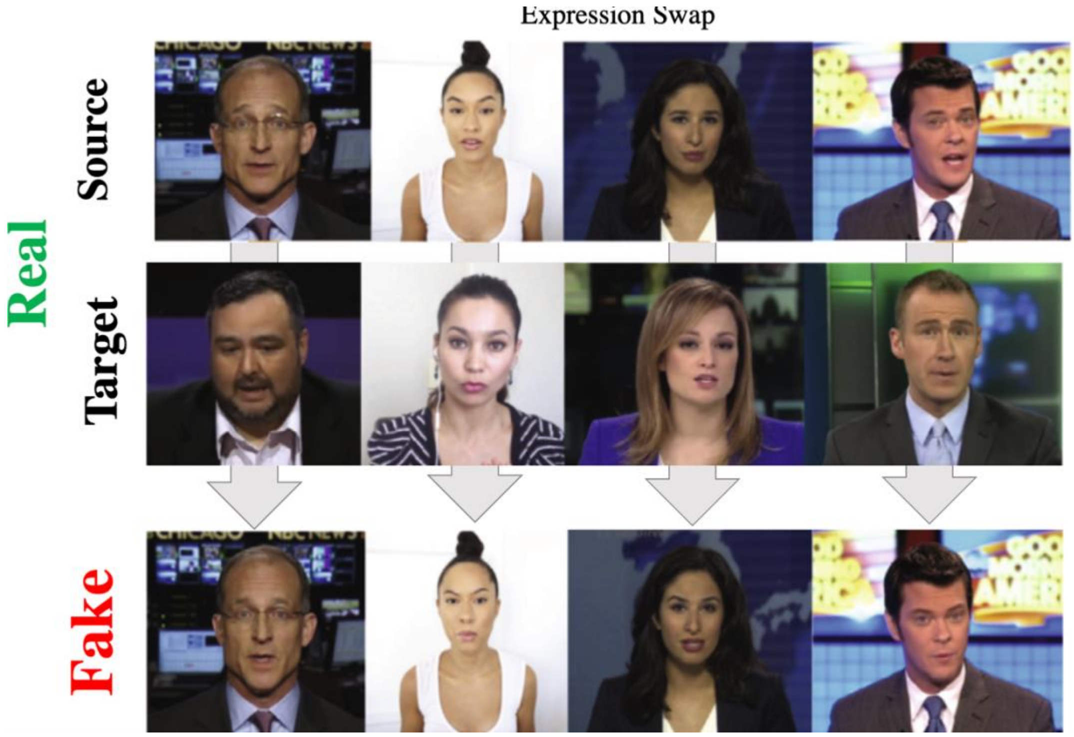
[193] Example groups are borrowed from Ruben Tolosana and others, 'Deepfakes and Beyond: A Survey of Face Manipulations and Fake Detection' (2020) 64 Information Fusion 131, 133.

[194] Real images retrieved from whichfaceisreal.com and the deepfake images from thispersondoesnotexist.com.

[195] Real images retrieved from whichfaseisreal.com, and the fake images generated by FaceApp (www.faceapp.com).



Identity swap shows how target face has been embedded on source material to generate new fake content.[196] Identity swap – also known as face swap – has been most prominent way of generating deepfake pornographic material.



Expression swap modifies the facial expression of the person depicted in source material.[197]

[196] Images are extracted from Celeb-DF (version 1) database (https://github.com/yuezunli/celeb-deepfakeforensics) accessed 22 August 2021.

[197] Images retrieved from FaceForensics database (https://github.com/ondyari/FaceForensics/tree/master/dataset) but these specific examples are borrowed from Ruben Tolosana and others, 'Deepfakes and beyond: A survey of face manipulations and fake detection' (2020) 64 Information Fusion 131, 133.

Appendix II – Use of deepfake technologies

Below rubric provides examples of different scenarios for illustrating how different deepfake technologies can be used in different situations. This exemplary chart is merely illustrative, and by no means provides a complete picture of the possibilities of deepfakes and deepfake technologies.

Target	Scenario	Role of Deepfake	Key Technique
Individuals	Identity thefts	Deepfake audio or deepfake video used to impersonate some trusted individual to gain personal information.	Voice cloning, voice phishing ('vishing') or face-swap videos
	Cyber Extortions	Deepfake pornographic image/video is used for blackmailing or causing distress.	Deepfake images or deepfake videos
	Vocal reproductions[198]	Deepfake audio used to make reproductions of songs or texts by using (famous) individuals' voices.	Deepfake audio, voice cloning
Society	Fabricated events	Deepfake audio, face-swap, or expression swap used to create a fabricated event to disseminate unauthentic information. Can be political or corporate leaders, or altered images of events (eg geographical images).	Deepfake audio, deepfake images, deepfake videos (both visual and audiovisual)
	Fraud Schemes[199]	Deepfake technology used on images (real images, gathered online, or bought on black market) to create synthetic identities and to spoof face recognition and liveness.	Deepfake images

[198] Vocal Synthesis, 'Jay-Z covers "We Didn't Start The Fire" by Billy Joel (Speech Synthesis)' (*YouTube*, 25 April 2020) <https://youtu.be/iyemXtkB-xk> accessed 28 August 2021.

[199] Masha Borak, 'Chinese Government-Run Facial Recognition System Hacked by Tax Fraudsters: Report' *South China Morning Post* (31 March 2021) <www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax> accessed 28 August 2021.

Appendix III – Detection Tools

Method	Core function	How it works	Detectable types
Reverse Engineering[200]	The method relies on uncovering the one-of-a-kind patterns behind the AI model utilised to generate a single deepfake image.	Deepfake image is run through a fingerprint estimation network (FEN) to analyse details of the fingerprint left by generative models. When generative models (eg GANs) are used for generating images, this reverse engineering model can identify the traces of the fingerprints and attribute the image where it came from.	Deepfake images
Attentional Visual Heartbeat[201]	The method judges whether the normal heart rates in videos are diminished. Videos are cut to become face videos.	The model analyses very small periodic changes of skin colour, which changes due to blood pumping through the face.	Deepfake videos
Optical Flows[202]	In this method, video frames are processed into square-shaped images that have optical flow fields placed over them.	The differences in optical flow fields are analysed and compared with a file that is known to be authentic to conclude whether the video is fake.	Deepfake videos
Convolutional Traces[203]	This method analyses images' unique traces and compares the traces within the images in the dataset.	This method investigates the images' convolutional traces (like breadcrumbs left from the generative tool) and attach the image to its generative architecture – or if real, identify the device (camera) used.	Deepfake images
Face Warping Artifacts[204]	This method analyses face areas of videos and the resolution inconsistencies due to face warping.	The method compares deepfake faces with a dedicated Convolutional Neural Network (CNN) model.	Deepfake videos

[200] Xi Yin and Tal Hassner, 'Reverse engineering generative models from a single deepfake image' (*Facebook AI,* 16 June 2021) <https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image> accessed 28 June 2021.

[201] Hua Qi and others, 'DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythm', in Proceedings of the 28th ACM International Conference on Multimedia (MM'20) (Association for Computing Machinery 2020) 4318.

[202] Roberto Caldelli and others, 'Optical Flow based CNN for detection of unlearnt deepfake manipulations' (2021) 146 Pattern Recognition Letters 31.

[203] Luca Guarnera and Sebastiano Battiato, 'Fighting Deepfake by Exposing the Convolutional Traces on Images' (IEEE Access 2020).

[204] Yuezun Li and Siwei Lyu, *Exposing DeepFake Videos By Detecting Face Warping Artifacts* (2018) available at <https://arxiv.org/abs/1811.00656v3> accessed 28 August 2021.

Appendix IV – Marking Methods

Method	What content	How it works	Example
Blockchain[205]	Audiovisual (most likely), visual	Add hash at data source, validate data at every stage, and recognise instances of video tampering.	Used already for image protection (copyright) reasons,[206] could be utilised for authentication.
Watermarking	Audiovisual, visual	Any deepfake generator applies automatically a watermark on the content that cannot be removed.	FaceApp applies a 'FaceApp' watermark on each picture generated within the application. Can be removed in the paid version
Labelling	Visual, audio, audiovisual	Online platforms can automatically add a label to the content that cannot be removed by the platform users.	Similar to Covid-19 labels applied by social media sites[207]
Faketagging[208]	Visual	The method applies an image tag (invisible for human eye, cf watermarks) to visual content through a five-step pipeline process. The FakeTagger can then recover the embedded tags from GAN-generated images, even if the particular GAN tool is unknown.	Sensible tagging has been used in food safety. Faketagging could protect personal image spreading, and tracing images' manipulation.

[205] Abbas Yazdinejad and others, 'Making Sense of Blockchain for Deepfake Technology' in *2020 IEEE Globecom Workshops* (IEEE 2020) 1; Hasan and Salah (n 117).

[206] For example, Binded provides this type of service <https://binded.com/>.

[207] Yoel Roth and Nick Pickles, 'Updating our Approach to Misleading Information' (*Twitter blog,* 11 May 2020) <https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information> accessed 26 August 2021, Elizabeth Culliford, 'Facebook to Label All Posts About COVID-19 Vaccines' *Reuters* (15 March 2021) <www.reuters.com/article/us-health-coronavirus-facebook-idUSKBN2B70NJ> accessed 26 August 2021.

[208] Run Wang and others, '*FakeTagger*: Robust Safeguards against DeepFake Dissemination via Provenance Tracking' (2021) available at: <https://arxiv.org/abs/2009.09869v2> accessed on 28 August 2021.

Bibliography

Legislation

Australian Taxation (Unpaid Company Tax) Assessment Act 1982

Californian Bill, AB-602 Depiction of individual using digital or electronic technology: sexually explicit material: cause of action

Californian Bill, AB-730 Elections: deceptive audio or visual media

Chinese Provisions on the Management of Online A/V Information Services (关于印发《网络音视频信息服务管理规定》的通知)

Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended)

Gambian Constitution of the Republic of Gambia

International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171

Israeli Nazis and Nazi Collaborators (Punishment) Law 5710–1950 (1949-1950) 4 LSI 154

New York Bill, S5959/A5605C

Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L 119

Russian Constitution of the Russian Federation

Swedish Constitution, Regeringsformen

Texas Bill S.B.751 Texas Capitol.

UK, Criminal Justice and Courts Act 2015

UK, Draft Online Safety Bill

UK, Finance Act 2006

UK, Human Rights Act 1998

UK, Locomotives Act 1861

UK, Voyeurism (Offences) Act 2019

UK, War Crimes Act 1991

Universal Declaration of Human Rights (adopted 10 December 1948 UNGA Res 217 A(III) (UDHR)

US Code Title 15, Section 1125

US Constitution

US Federal Rules of Evidence Rule 901(b)(9)

US, H.R. 3230, 116th Congress (2019-2020) Defending Each and Every Person from False appearances by Keeping Exploitation Subject to Accountability Act of 2019 (DEEP FAKES Accountability Act)

US, H.R. 4355, 116th Congress (2019-2020) Identifying Outputs of Generative Adversarial Networks Act (IOGAN)

US, H.R.6088, 116th Congress (2019-2020) Deepfakes in Federal Elections Prohibition Act.

US, H.R. 6395, 116th Congress (2019-2020) National Defense Authorization Act for Fiscal Year (NDAA) 2021

US, S. 1348, 116th Congress (2019-2020) A bill to require the Secretary of Defense to conduct a study on cyberexploitation of members of the Armed Forces and their families, and for other purposes.'

US, S. 2065, 116th Congress (2019-2020) Deepfake Report Act of 2019

US, S. 2559, 117th Congress (2021-2022) Deepfake Task Force Act

Vienna Convention on the Law of Treaties, opened for signature (23 May 1969) 115 U.N.T.S. 331

Virginian Code, Unlawful dissemination or sale of images of another; penalty, s. 18.2-386.2

Cases

ECtHR

Ahmet Yıldırım v Turkey App no 3111/10 (ECtHR, 18 December 2012)

Aleksey Ovchinnikov v Russia App no 24061/04 (ECtHR, 16 December 2010)

Delfi AS v Estonia App no 64569/09 (ECtHR, 10 October 2013)

Féret v Belgium App no 15615/07 (ECtHR, 16 July 2009)

GRA Stiftung Gegen Rassismun und Antisemitismus v Switzerland App no 18597/13 (ECtHR, 9 January 2018

Handyside v The United Kingdom App no 5493/72 (ECtHR, 7 December 1976)

Magyar Tartalomszolgáltatók Egyesülete and Index.Hu ZRT v Hungary App no 22947/13 (ECtHR, 5 January 2016)

Renaud v France App no 13290/07 (ECtHR, 25 February 2010)

von Hannover v Germany (no.2) App no 40660/08 (ECtHR, 7 February 2012)

Willem v France App no 10883/05 (ECtHR, 16 July 2009)

US

Abdul-Jabbar v General Motors Corp (1996) 85 F 3d 407

Cher v Forum International Ltd (1982) 213 USPQ 96 (CD Cal)

Eastwood v Superior Court (National Enquirer Inc) (1983) 149 Cal. App. 3d 409

Griswold v Connecticut (1965) 381 U.S. 479

Haelan Laboratories, Inc v Topps Chewing Gum, Inc, (1953) 202 F. 2D 866 cert. denied 346 US 816, 98L. Ed. 343, 74 S. Ct. 26 (2nd Cir)

McIntyre v Ohio Elections Commission (1995) 514 U.S. 344.

re NCAA Student-Athlete Name & Likeness Licensing Litigation (2013) 724 F.3d 1268, 1279 (9th Cir)

Onassis v Christian Dior New York, Inc, (1984) 122 Misc 2d 603, 427 NYS2d 254

United States v. Alvarez (2012) 567 U.S. 709

United States v. Gagliardi (2007) 506 F.3d 140, 151 (2nd Cir)

Waits v Frito Lay, Inc, (1992) 978 F 2d 1093 (9th Cir).

White v Samsung Electronics America, Inc, (1992) 971 F2d 1395.

UK

Beswick v Beswick [1967] UKHL 2, [1968] AC 58

Lachaux v Independent Print Ltd, [2019] UKSC 27

R (on the application of ProLife Alliance) v. British Broadcasting Company [2003] UKHL 23, [2004] 1 A.C. 185

Other

C-201/13 Deckmyn and Vrijheidsfonds VZW v Helena Vandersteen [2014] ECLI:EU:C:2014:2132

Cases C-244/10 and C-245/10 Mesopotamia Broadcast A/S METV (C-244/10) and Roj TV A/S (C-245/10) v Bundesrepublik Deutschland [2011] ECR I-8777

Secondary Sourers

– –, 'Art term – Cubism' (*Tate.org)* <www.tate.org.uk/art/art-terms/c/cubism>

– –, 'The Ethics Of Deepfake Aren't Always Black And White' (*The Next Web,* 16 June 2019) <https://thenextweb.com/news/the-ethics-of-deepfakes-arent-always-black-and-white>

– –, 'Healthcare innovation main driver of European patent applications in 2020' (*epo.org*, 16 March 2021) <www.epo.org/news-events/news/2021/20210316.html>

– –, 'MIT Media Lab' <www.media.mit.edu/users/login/?next=%2Fcourses%2Fthe-ethics-and-governance-of-artificial-intelligence%2F>

– –, 'Not Without My Consent: The Facebook Pilot' (*Revenge Porn Helpline*) <https://revengepornhelpline.org.uk/information-and-advice/reporting-content/facebook-pilot/>

– –, 'Operationalizing responsible AI' (*Microsoft.com*) <www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5>

– –, 'Responsible AI Practices' (*Google AI*) <https://ai.google/responsibilities/responsible-ai-practices/>

Adjer H, and others, *The State of Deepfakes 2019 Landscape, Threats, and Impact* (Deeptrace 2019)

Adjer H, and Schick N, 'Deepfake apps are here and we can't let them run amok' *Wired* (30 March 2021) <www.wired.co.uk/article/deepfakes-security>

Agarwal S, and others, 'Protecting World Leaders Against Deep Fakes' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (IEEE 2019)

AISG Student Chapter, 'Deepfakes & Misinformation: The Ethics Behind AI' AI Singapore (1 June 2021) <https://aisingapore.org/2021/06/deepfakes-misinformation-the-ethics-behind-the-ai/>

Albright T, 'Why eyewitness fail' (2017) 114(30) Proceedings of National Academy of Sciences 7758

Alter A, *Irresistible: The Rise of Addictive Technology and the Business of Keeping Us Hooked* (Penguin Books 2018)

Awah Buo S, 'The Emerging Threats of Deepfake Attacks and Countermeasures' (2020 http://dx.doi.org/10.13140/RG.2.2.23089.81762)

Ayyub R, 'I was the Victim of a Deepfake Porn Plot Intended to Silence Me' (*HuffPost the Blog,* 21 November 2018) <www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316>

Balenciaga, 'Spring 22 Collection Note' (*balenciaga.com*) <www.balenciaga.com/en-gb/spring-22>

Barendt E, 'Freedom of Expression in the United Kingdom Under the Human Rights Act 1998' (2009) 84(3) Indiana Law Journal 851

Bateman J, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios' (*Carnegie Endowment for International Peace,* 8 July 2020) <https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237>

BCS, 'Social media must verify users' ID to end online abuse - IT industry poll' (*BCS.org,* 29 April 2021) <www.bcs.org/more/about-us/press-office/press-releases/social-media-must-verify-users-id-to-end-online-abuse-it-industry-poll/>

Beres D, and Gilmer M, 'A Guide to 'Deepfakes,' the Internet's Latest Moral Crisis' *Mashable* (2 February 2018) <https://mashable.com/2018/02/02/what-are-deepfakes/#pNi2cZMBtqqM>

Bodle R, 'The ethics of online anonymity or Zuckerberg vs "moot"'2013) 43(1) Computer and Society 22

Borak M, 'Chinese Government-Run Facial Recognition System Hacked by Tax Fraudsters: Report' *South China Morning Post* (31 March 2021) <www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax>

Bowman E, 'Slick Tom Cruise Deepfakes Signal That Near Flawless Forgeries May Be Here' *npr* (11 March 2021) <www.npr.org/2021/03/11/975849508/slick-tom-cruise-deepfakes-signal-that-near-flawless-forgeries-may-be-here>

Brady M, 'Deepfakes: a New Disinformation Threat?' *Democracy Reporting International,* August 2020) <https://democracy-reporting.org/wp-content/uploads/2020/08/2020-09-01-DRI-deepfake-publication-no-1.pdf>

Buckhout R, Figueroa D and Hoff E, 'Eyewitness identification: Effects of suggestion and bias in identification from photographs' (1975) 6(1) Bulletin of the Psychonomic Society 71

Burges M, 'A Deepfake Porn Bot is Being Used to Abuse Thousands of Women' *Wired* (20 October 2020) <www.wired.co.uk/article/telegram-deepfakes-deepnude-ai>

BuzzFeedVideo, 'You Won't Believe What Obama Says In This Video!' (*YouTube,* 17 April 2018) <www.youtube.com/watch?v=cQ54GDm1eL0>

Cahlan S, ‘How Misinformation Helped Spark an Attempted Coup in Gabon’ *The Washington Post* (13 February 2020) <www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/>

Caldelli R, and others, ’Optical Flow based CNN for detection of unlearnt deepfake manipulations’ (2021) 146 Pattern Recognition Letters 31

Caldera E, ’”Reject The evidence of your eyes and ears”: deepfake and the law of virtual replicants' (2019) 50 Seton Hall Law Review 177

Campbell H, and Holroyd M, ''Deepfake geography' could be the latest form of online disinformation' *Euronews* (7 May 2021) <www.euronews.com/2021/05/07/deepfake-geography-could-be-the-latest-form-of-online-disinformation>

Chen J, ‘Deepfakes’ (The Asia Society 2020) <https://asiasociety.org/sites/default/files/inline-files/Final%20Deepfake%20PDF.pdf>

Chesney R, and Citron D, 'Deepfakes and the New Information War' (2019) 98(1) Foreign Affairs 147

Citron D, ‘Sexual Privacy’ (2019) 28 Yale Law Journal 1870
– – and Chesney R, ’Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security’ (2019) 107 California Law Review 1753

Clark K, and Moore N, ‘Buying Voice: Financial Rewards for Whistleblowing Lawyers’ (2015) 56(5) Boston College Law Review 169

Çolak B, 'Legal Issues of Deepfakes' (*Institute for Internet & the Just Society*, 19 January 2021) <www.internetjustsociety.org/legal-issues-of-deepfakes>

Colander N, and Quinn M, 'Deepfakes and the Value-Neutrality Thesis' (*SeattleU*, 10 February 2020) <www.seattleu.edu/ethics-and-technology/viewpoints/deepfakes-and-the-value-neutrality-thesis.html>

Cole S, ’AI-Assisted Fake Porn is Here and We’re All Fucked’ *Motherboard* (11 December 2017) <www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn>
– – ’Pornhub Is Banning AI-Generated Fake Porn Videos, Says They’re Nonconsensual’ *Motherboard* (6 February 2018) <www.vice.com/en/article/zmwvdw/pornhub-bans-deepfakes>
– – ‘Twitter Is the Latest Platform to Ban AI-Generated Porn’ *Motherboard* (7 February 2018) <www.vice.com/en/article/ywqgab/twitter-bans-deepfakes>
– – ‘The Ugly Truth Behind Pornhub’s “Year In Review”’ *Vice* (18 February 2020) <www.vice.com/en_us/article/wxez8y/pornhub-year-in-review-deepfake>

Commission, ‘Proposal for a regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative acts’ COM (2021) 206 final

Committee on Legal Affairs, ’Report on intellectual property rights for the development of artificial intelligence technologies’ 2020/2015INI (2 October 2020) <www.europarl.europa.eu/doceo/document/A-9-2020-0176_EN.html>

Corcoran M, and Henry M, 'The Tom Cruise deepfake that set off 'terror' in the heart of Washington DC' *ABC News* (24 June 2021) <https://www.abc.net.au/news/2021-06-24/tom-cruise-deepfake-chris-ume-security-washington-dc/100234772>

Cram I, *Contested Words: Legal Restrictions on Freedom of Speech in Liberal Democracies* (Routledge 2016)

Cruz S, 'Legal Opinion on Deepfakes and Platform Responsibility' (LLM Thesis, King's College London 2020)

Culliford E, ‘Facebook to Label All Posts About COVID-19 Vaccines’ *Reuters* (15 March 2021) <www.reuters.com/article/us-health-coronavirus-facebook-idUSKBN2B70NJ>

Cutler B, Penrod S and Martens T, ‘Reliability of Eyewitness Identification the Role of System and Estimator Variables’ (1987) 11(3) Law and Human Behavior 233

Daniel C, and O’Flaherty A, ‘The Rise of the “Deepfake” Demands Urgent Legal Reform in the UK’ (2021) XI(82) National Law Review

Davis R, Wiggins C, and Donovan J, 'Technology Factsheet: Deepfakes' in Jayanti A (ed) *Tech Factsheets for Policymakers: Deepfakes* (Harvard College 2020)

de Ruiter A, ‘The Distinct Wrong of Deepfakes’ (2021) Philosophy & Technology <https://link.springer.com/article/10.1007/s13347-021-00459-2>

De Saulles M, 'How deepfakes are a problem for us all and why the law needs to change' *Information matters* (26 March 2021) <https://informationmatters.net/deepfakes-problem-why-law-needs-to-change/>

Debusmann B, 'Deepfake is the Future of Content Creation' *BBC* (8 March 2021) <www.bbc.co.uk/news/business-56278411>

Diakopoulos N, and Johnson D, ‘Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections’ (2020) 23(7) New Media & Society 2072

Diaz J, 'Facebook Researchers Say They Can Detect Deepfakes and Where They Came From' *npr* (17 June 2021) <www.npr.org/2021/06/17/1007472092/facebook-researchers-say-they-can-detect-deepfakes-and-where-they-came-from?t=1625236130581>

Dougherty W, 'Deepfake deception: the emerging threat of deepfake attacks' (*Dentons,* 21 May 2021) <www.dentons.com/en/insights/articles/2021/may/21/deepfake-deception-the-emerging-threat-of-deepfake-attacks>

Drinnon C, 'When Fame Takes Away the Right to Privacy in One's Body: Revenge Porn and Tort Remedies for Public Figures' (2017) 24(1) William & Mary Journal of Race, Gender, and Social Justice 209

Dufour N, and Gully A, 'Contributing Data to Deepfake Detection Research' (*Google AI blog,* 24 September 2019) <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>

Ehrenkranz M, 'When You Can Make 'JFK' Say Anything, What's Stopping Him From Selling Doritos?' (*Gizmodo*, 16 March 2018) <https://gizmodo.com/when-you-can-make-jfk-say-anything-whats-stopping-him-1823820892>

Europol, 'Malicious Uses and Abuses of Artificial Intelligence' (2020 Trend Micro Research)
Fagni T, and others, 'TweepFake: About detecting deepfake tweets' (*PLOS ONE,* 13 May 2021) available at <https://doi.org/10.1371/journal.pone.0251415>

Farish K, ’Do Deepfakes Pose a Golden Opportunity? Considering Whether English Law Should Adopt California’s Publicity Right in the Age of the Deepfake’ (2020) 15 Journal of Intellectual Property Law and Practice 40
– – 'The legal implications and challenges of deepfakes' (*Dac Beachcroft,* 4 September 2020) <www.dacbeachcroft.com/en/gb/articles/2020/september/the-legal-implications-and-challenges-of-deepfakes/> accessed 30 June 2021

Feeney M, ‘Deepfake Laws Risk Creating More Problems than They Solve’ (*Regulatory Transparency Project*, 1 March 2021) <https://regproject.org/wp-content/uploads/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf>

Ferraro M, Chipman J, and Preston S, 'The Federal "Deepfakes" Law' (2020) 3(4) The Journal of Robotics, Artificial Intelligence & Law 229

Fletcher J, ‘Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-fact Performance’ (2018) 70 Theatre Journal 455

Franklin S, ‘This Bill Hader Deepfake Video Is Amazing. It's Also Terrifying for Our Future’ *New America* (13 August 2019) <www.newamerica.org/oti/in-the-news/bill-hader-deepfake-video-amazing-its-also-terrifying-our-future/>

Galston W, 'Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics' *Brookings* (8 January 2020) <www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/#cancel>

Gardner, E, 'Deepfakes Pose Increasing Legal and Ethical Issues for Hollywood' *The Hollywood Reporter* (12 July 2019) <www.hollywoodreporter.com/business/business-news/deepfakes-pose-increasing-legal-ethical-issues-hollywood-1222978/>
– – ’Disney Comes Out Against New York’s Proposal to Curb Pornographic “Deepfakes”’ *The Hollywood Reporter* (11 June 2018) <www.hollywoodreporter.com/business/business-news/disney-new-yorks-proposal-curb-pornographic-deepfakes-1119170/>

Gerstner E, ’Face/Off: ”DeepFake” Face Swaps and Privacy Laws’ (2020) 87(1) Defense Counsel Journal 1

Ghidini G, 'Legal Opinion: Deepfakes and the Collapse of Truth: a Comparative Analysis between American and European Reactions to a New Systemic Threat' (LLM Thesis, King's College London 2020)

Giardina C, 'How "Furious 7" Brought the Late Paul Walker Back to Life' *The Hollywood Reporter* (11 December 2015) <www.hollywoodreporter.com/behind-screen/how-furious-7-brought-late-845763>

Goodfellow I, and others, 'Generative Adversarial Networks' (2014) 2 Advances in Neural Information Processing Systems 2672

Goodwine K, 'Ethical Considerations of Deepfakes' *The Prindle Post* (7 December 2020) <www.prindlepost.org/2020/12/ethical-considerations-of-deepfakes/>

Green A, 'Lawmakers and Tech Groups Fight Back Against Deepfakes' *Financial Times* (30 October 2019) <www.ft.com/content/b7c78624-ca57-11e9-af46-b09e8bfe60c0>

Greene D, 'We Don't Need New Laws for Faked Videos, We Already Have Them' (*Electronic Frontier Foundation* 13 February 2018) <www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them>

Gregory S, 'Authoritarian Regimes Could Exploit Cries of 'Deepfake'' *Wired* (14 February 2021) <www.wired.com/story/opinion-authoritarian-regimes-could-exploit-cries-of-deepfake/>

Guarnera L, and Battiato S, 'Fighting Deepfake by Exposing the Convolutional Traces on Images' (IEEE Access 2020)

Güera D, and Delp E, 'Deepfake Video Detection Using Recurrent Neural Networks' (15th IEEE International Conference on Advanced Video and Signal Based Surveillance 2018) available at <https://ieeexplore.ieee.org/document/8639163>

Gurry F, 'Artificial intelligence and intellectual property: An interview with Francis Gurry' 'Artificial intelligence and intellectual property: An interview with Francis Gurry' (2018) 114 Intellectual Property Forum: Journal of the Intellectual and Industrial Property Society of Australia and New Zealand 66

Hall H, 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27(1) Catholic University Journal of Law and Technology 51

Harris D, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You' (2019) 7 Duke Law and Technology Review 99

Hasan H, and Salah K, 'Combating Deepfake Videos using Blockchain and Smart Contracts' (2019) 7 IEEE Access 41596

Hellerman J, 'Tom Cruise Deepfake Raises Ethics Concerns' (*No Film School*, 8 March 2021) <https://nofilmschool.com/tom-cruise-deep-fake>

Hilty R, Hoffmann J, and Scheuerer S, 'Intellectual Property Justification for Artificial Intelligence' in Lee J-A, Liu K-C, Hilty R (eds), *Artificial Intelligence & Intellectual Property* (OUP 2020)

HM Government, *Regulation for the Fourth Industrial Revolution, White Paper* (HM Government 2019)

Horton A, 'The Bourdain AI furore shouldn't overshadow an effective, complicated film' *The Guardian* (21 July 2021) <www.theguardian.com/film/2021/jul/21/anthony-bourdain-documentary-roadrunner-ai-deepfake-furore>

House of Commons Library, 'Online Anonymity and Anonymous Abuse' (*Research Briefing, UK Parliament,* 23 March 2021) <https://commonslibrary.parliament.uk/research-briefings/cdp-2021-0046/>

Hu J, and others, 'Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network' (IEEE Transaction on Circuits and Systems for Video Technology 2021) available at <doi:10.1109/TCSVT.2021.3074259>

Hulu, 'Hulu Has Live Sports: The Deepfake Hulu Commercial' (*YouTube*, 11 September 2020) <www.youtube.com/watch?v=yPCnVeiQsUw>

Hutchinson A, 'TikTok Moves to Furhter Limit Potential Exposure To Harmful Content Through Automated Removals' (*Social Media Today,* 9 July 2021) <www.socialmediatoday.com/news/tiktok-moves-to-further-limit-potential-exposure-to-harmful-content-through/603129/>

Hwang Y, Ryu J, and Jeong S, 'Effects of Disinformation Using Deepfake: The Protective effect of Media Literacy Education' (2021) 24(3) Cyberpsychology, Behavior, and Social Networking 188

Iacobucci S, and others, 'Deepfakes Unmasked: The Effects of Informatiton Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention' (2021) 23(3) Cyperpsychology, Behavior, and Social Networking 194

Ice J, 'Defamatory Political Deepfakes and the First Amendment' (2019) 70(2) Case Western Reserve Law Review 417

Inglesh A, 'Deepfakes May be in Deep Trouble: How the Law Has and Should Respond to the Rise of the AI-Assisted Technology of Deepfake Videos' (*Cardozo AELJ*, 19 January 2020) <https://cardozoaelj.com/2020/01/19/deepfakes-deep-trouble/>

Instagram, ‘Copyright’ (*Instagram Help Centre*) <https://help.instagram.com/126382350847838>

Jackson E, 'Face swap: France's Top Soap Uses 'Deepfake' Technology for Self-Isolating Actress' *France24* (10 December 2020) <www.france24.com/en/tv-shows/encore/20201210-face-swap-france-s-top-soap-uses-deepfake-technology-for-self-isolating-actress>

Jacoby E, 'I Paid $30 to Create a Deepfake Porn of Myself' *Motherboard* (9 December 2019) <www.vice.com/en/article/vb55p8/i-paid-dollar30-to-create-a-deepfake-porn-of-myself>

Jaiman A, 'Debating the ethics of deepfakes' (Observer Research Foundation 27 August 2020) <www.orfonline.org/expert-speak/debating-the-ethics-of-deepfakes/>
– – 'deepfakes though an ethical lense' (*Ashish Jaiman,* 2 August 2020) <https://ashishjaiman.medium.com/deepfakes-though-an-ethical-lens-bd4301b41e52>

Jain S, and Jha P, 'Deepfakes in India: Regulation and Privacy' (*LSE Blogs*, 21 May 2020) <https://blogs.lse.ac.uk/southasia/2020/05/21/deepfakes-in-india-regulation-and-privacy/>

James B, 'Why You and The Court Should Not Accept Audio or Video Evidence at Face Value: How Deepfake Can Be Used to Manufacture Very Plausible Evidence' (2020) 43 International Family Law 41

Jha J, 'Internet Censorship: The Ethical Dimension' (IBS Research Center 2007)

Jing M, 'China Issues New Rules to Clamp Down on Deepfake Technologies Used to Create and Broadcast Fake News' *South China Morning Post* (29 November 2019) <www.scmp.com/tech/apps-social/article/3039978/china-issues-new-rules-clamp-down-deepfake-technologies-used?module=perpetual_scroll&pgtype=article&campaign=3039978>

Johnson D, and Diakopoulos N, 'What To Do About Deepfakes' (2021) 64(3) Communications of the ACM 33

Jones A, 'Assessing the Real Threat Posed by Deepfake Technology' (*International Banker,* 24 February 2021) <https://internationalbanker.com/technology/assessing-the-real-threat-posed-by-deepfake-technology/>

Karasavva V, and Noorbhai A, 'The Real Threat of Deepfake Pornography: A Review of Canadian Policy' (2021) 24(3) Cyberpsychology, Behavior, and Social Networking 203

Kesvani H, 'Abolishing Online Anonymity Won't Tackle the Underlying Problems of Racist Abuse' *The Guardian* (15 July 2021) <www.theguardian.com/commentisfree/2021/jul/15/abolishing-online-anonymity-racist-abuse-id-verification>

Kietzmann J, and others, 'Deepfakes: Trick or Treat?' (2020) 63(2) Business Horizons 135

Kokane S, 'The Intellectual Property Rights of Artificial Intelligence-based Inventions' (2021) 65(2) Journal of Scientific Research 116

Kryvoi Y, and Matos S, 'Non-Retroactivity as a General Principle of Law', (2021) 17(1) Utrecht Law Review 46

Laney R, emails to the author (13 March 2021 – 12 July 2021)

Langa J, 'Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes' (2021) 101(2) Boston University Law Review 761

Leval P, ‘Towards a Fair Use Standard’ (1990) 103(5) Harvard Law Review 1105

Li Y, and Lyu S, ‘Exposing DeepFake Videos By Detecting Face Warping Artifacts’ (arXiv2018) available at <https://arxiv.org/abs/1811.00656v3>

Li Y, and others, 'DeepFake-o-meter: An Open Platform for Deepfake Detection' (arXiv 2021) available at <arXiv:2103.02018>

Lindsay R, and Wells G, ‘Improving Eyewitness Identifications From Lineups: Simultaneous Versus Sequential Lineup Presentation’ (1985) 70(3) Journal of Applied Psychology 556

Lomas N, ‘Duplex Shows Google Failing at Ethical and Creative AI Design’ (*TechCrunch*, 10 May 2018) <https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/>

Mackintosh E, ’Finland is winning the war on fake news. What it’s learned may be crucial to Western democracy’ *CNN* (May 2019) <https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/>

Makeen M, ‘Rationalising performance “in public” under UK copyright law’ (2016) 2 Intellectual Property Quarterly 117

Malaria Must Die, 'David Beckham Speaks Nine Languages to Launch Malaria Must Die Voice Petition' (*YouTube*, 9 April 2019) <www.youtube.com/watch?v=QiiSAvKJIHo>
– –, 'A World Without Malaria' (*YouTube,* 3 December 2020) <www.youtube.com/watch?v=0l4eTfpIsKw>

Marcum T, Young J, and Kirner E, ‘Blowing the Whistle in the Digital Age: Are You Really Anonymous? The Perils and Pitfalls of Anonymity in Whistleblowing Law’ (2020) 17(1) DePaul Business and Commercial Law Journal 1

Meskys E, and others, 'Regulating Deep-Fakes: Legal and Ethical Considerations' (*SSRN,* 2 December 2019) available at <https://ssrn.com/abstract=3497144>

Metz C, ’Internet Companies Prepare to Fight the ‘Deepfake’ Future’ *The New York Times* (24 November 2019) <www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html>

Miller J, 'Is Social Media Censorship Legal?' *The Echo* (22 February 2021) <www.theechonews.com/article/2021/02/kdyntyejlxplpeh>

Miller K, 'Advantages and Disadvantages Of Internet Censorship' (Future of Working) <https://futureofworking.com/8-advantages-and-disadvantages-of-internet-censorship/>

Minna, 'Deepfakes: An Unknown and Unchartered Legal Landscape' (*towards data science,* 17 July 2019) <https://towardsdatascience.com/deepfakes-an-unknown-and-uncharted-legal-landscape-faec3b092eaf>

Morawetz N, 'Determining the Retroactive Effect of Laws Altering the Consequences of Criminal Convictions' (2003) 30(5) Fordham Urban Law Journal 1743

Mostert F, and Franks H, 'How to Counter Deepfakery in the Eye of the Digital Deceiver' *Financial Times* (18 June 2020) <www.ft.com/content/ea85476e-a665-11ea-92e2-cbd9b7e28ee6>

National Program for Artificial Intelligence, 'Deepfake Guide July 2021' (*www.ai.gov.ae* 2021) <https://ai.gov.ae/wp-content/uploads/2021/07/AI-DeepFake-Guide-EN-2021.pdf>

Neethirajan S, *Beyond Deepfake Technology Fear: On its Positive Uses for Livestock Farming* (Preprints 2021) available at <doi:10.20944/preprints202107.0326.v1>

Öhman C, 'Introducing the Pervert's Dilemma: a Contribution to the Critique of Deepfake Pornography' (2020) 22 Ethics and Information Technology 133

Owen-Jackson C, 'What does the Rise of Deepfakes Mean for the Future of Cybersecurity?' (*Kapersky,* 2019), <www.kaspersky.com/blog/secure-futures-magazine/deepfakes-2019/28954/>

Panetta F, and Burgund H, *In Event of Moon Disaster* (MIT 2019) <https://moondisaster.org>

Paris B, and Donovan J, 'Deepfakes and Cheapfakes' (*Data&Society* 18 September 2019) <https://datasociety.net/library/deepfakes-and-cheap-fakes/>

Pearlman R, 'Recognizing Artificial Intelligence (AI) as Authors and Inventors under U.S. Intellectual Property Law' (2018) 24(2) Richmond Journal of Law & Technology 1

Peckham O, 'New AI Model From Facebook, Michigan State Detects & Attributes Deepfakes' (*Datanami*) <www.datanami.com/2021/06/25/new-ai-model-from-facebook-michigan-state-detects-attributes-deepfakes/>

Perot E, and Mostert F, 'Fake it till You Make it: An Examination of the US and English Approaches to Persona Protection as Applied to Deepfakes on Social Media' (2020) 15 Journal of Intellectual Property Law and Practice 32

Pepper C, and others, 'Reputation Management and the Growing Threat of Deepfakes' (*Bloomberg Law,* <https://news.bloomberglaw.com/tech-and-telecom-law/reputation-management-and-the-growing-threat-of-deepfakes>)

Pfefferkorn R, '"Deepfakes" in the Courtroom' (2020) 29(2) Boston University Law Journal 245

Pitt J, 'Deepfake Videos and DDoS Attacks (Deliberate Denial of Satire)' (IEEE Technology and Society Magazine 2019)

Poetker B, 'How Internet Censorship Affects You (+Pros & Cons)' (*G2*, 18 November 2019) <www.g2.com/articles/internet-censorship>

Portuese A, Gough O and Tanega J, 'The Principle of Legal Certainty as a Principle of Economic Efficiency' (2017) 44 European Journal of Law and Economics 13

Practical Law, ’Limitation Periods’ (*Thomson Reuters Practical Law*, 1 May 2021) <https://uk.practicallaw.thomsonreuters.com/1-518-8770?transitionType=Default&contextData=(sc.Default)&firstPage=true>

Qi H, and others, 'DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythm', in *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)* (Association for Computing Machinery 2020) 4318

Quach K, 'New York State is trying to ban 'deepfakes' and Hollywood isn't happy' *The Register* (12 June 2018) <www.theregister.com/2018/06/12/new_york_state_is_trying_to_ban_deepfakes_and_hollywood_isnt_happy/>

Quito A, 'The Anthony Bourdain audio deepfake is forcing a debate about AI in journalism' *Quartz* (18 July 2021) <https://qz.com/2034784/the-anthony-bourdain-documentary-and-the-ethics-of-audio-deepfakes/>

Rainie L, Anderson J, and Albright J, 'The Future of Free Speech, Trolls, Anonymity and Fake News Online' (*Pew Research Center,* 29 March 2017) <www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/>

Ritman A, ‘James Dean Reborn in CGI for Vietnam War Action-Drama (Exclusive)’ *The Hollywood Reporter* (6 November 2019) <www.hollywoodreporter.com/news/afm-james-dean-reborn-cgi-vietnam-war-action-drama-1252703>

Rosner H, 'The Ethics of a Deepfake Anthony Bourdain Voice' *The New Yorker* (17 July 2021) <www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice>

Rössler A, and others, ‘FaceForensics++: Learning to Detect Manipulated Facial Images’ (arXiv 2019) available at: <https://arxiv.org/abs/1901.08971>

Roth Y, and Pickles N, ’Updating our Approach to Misleading Information’ (*Twitter blog,* 11 May 2020) <https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information>

Rothkopf J, 'Deepfake Technology Enters the Documentary World' *The New York Times* (1 July 2020) <https://nyti.ms/31LSel5>

Royle S, ’’Deepfake porn images still give me nightmares’’ *BBC* (6 January 2021) <www.bbc.com/news/technology-55546372>

Sch_ B, 'Ethical Deepfakes' (*Bloom AI,* 22 December 2020) <https://medium.com/bloom-ai-blog/ethical-deepfakes-79e2e9eafad>

Schick N, *Deepfakes and the Infocalypse* (Monoray 2020)

Scott D, ‘Deepfake Porn Nearly Ruined My Life’ *Elle UK* (6 February 2020) <www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/>

SecurityInfoWatch, ‘U.S. intel agencies warn about deepfake video scourge' (*SecurityInfoWatch.Com* 2018) <www.proquest.com/trade-journals/u-s-intel-agencies-warn-about-deepfake-video/docview/2076909908/se-2?accountid=11862>

Sibley J, and Hartzog W, ‘The Upside of Deep Fakes’ (2019) 78(4) Maryland Law Review 960

Simonite T, ‘Deepfakes Are Now Making Business Pitches’ *Wired* (16 August 2021) <www.wired.com/story/deepfakes-making-business-pitches/>

Sjouwerman S, 'The Evolution of Deepfakes: Fighting the Next Big Threat' (*TechBeacon*) <https://techbeacon.com/security/evolution-deepfakes-fighting-next-big-threat>

Sloan G, 'Deepfake it 'Til You Make it: A Potentially Sinister Technology Also Has Role to Play in Marketing' (2019) 90(20) Advertising Age 15

Sorin V, and others, ‘Creating Artificial Images for Radiology Applications Using Generative Adversarial Networks (GANs) – A Systematic Review’ (2020) 27(8) Academic Radiology 1175.

Spivak R, ’”Deepfakes”: The Newest Way to Commit One of the Oldest Crimes’ (2019) 2 Georgetown Law Technology Review 339

State Farm Insurance, 'Predictions State Farm + ESPN Commercial (featuring Kenny Mayne)' (*YouTube*, 20 April 2020) <www.youtube.com/watch?v=FzOVqClci_s>

Stolton S, 'EU police recommend new online 'screening tech' to catch deepfakes' (*Euractive*, 20 November 2020) <www.euractiv.com/section/digital/news/eu-police-recommend-new-online-screening-tech-to-catch-deepfakes/>

Stupp C, ’Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case’ *Wall Street Journal* (30 August 2019) <www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402>

Sunstein C, ‘Falsehood and the First Amendment’ (2020) 33(2) Harvard Journal of Law and Technology 387

Tawepoon S, and Ardbroriak P, 'Challenges of Future Intellectual Property Issues For Artificial Intelligence' (*Tilleke & Gibbins,* 6 December 2018) <www.tilleke.com/insights/challenges-future-intellectual-property-issues-artificial-intelligence/>

Taylor B, 'Defending the State From Digital Deceit: The Reflexive Sexuritization of Deepfake' (2021) 38(1) Critical Studies in Media Communication 1

The Dalí Museum, 'Dali Lives - Art Meets Artificial Intelligence' (*YouTube*, 23 January 2019) <www.youtube.com/watch?v=Okq9zabY8rI>

The Next Web, 'The ethics od deepfake aren't always black and white' (16 June 2019) <https://thenextweb.com/news/the-ethics-of-deepfakes-arent-always-black-and-white>

The World Bank, ‘ID4D Data: Global Identification Challenge by the Numbers’ (*The World Bank Data Catalog,* 2018) <https://datacatalog.worldbank.org/dataset/identification-development-global-dataset>

Thomas R, 'AI Ethics Resources' (*Fast.ai,* 24 September 2018) <www.fast.ai/2018/09/24/ai-ethics-resources/>

Toews R, 'Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared' *Forbes* (25 May 2020) <www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/>

Tolosana R, and others, ‘Deepfakes and beyond: A survey of face manipulations and fake detection’ (2020) 64 Information Fusion 131

Tom, @DeepTomCruise (*TikTok*) <www.tiktok.com/@deeptomcruise?lang=en>

Uddin Mahmud B, and Sharmin A, 'Deep Insights of Deepfake Technology : A Review' (2020) 5(1&2) Dhaka University Journal of Applied Science and Engineering 13

UK Law Commission, ‘Automated Vehicles’ (*Law Commission*) <www.lawcom.gov.uk/project/automated-vehicles/#related>

UNGA, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue’ UN Doc A/HRC/17/27 (2011)

Vaccari C, and Chadwick A, ‘Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News’ (2020) Social Media + Society 6

van Huijstee M, and others, *Tackling Deepfakes in European Policy* (European Parliament Research Service STOA 2021)

Vavra S, ’Deepfake Laws Emerge as Harassment, Security Threats Come Into Focus’ (*Cyberscoop,* 11 January 2021) <www.cyberscoop.com/deepfake-porn-laws-election-disinformation/>

Venkataramakrishnan S, 'Behind the Tom Cruise deepfakes that can evade disinformation tools' *Financial Times* (5 March 2021) <www.ft.com/content/721da1df-a1e5-4e2f-97fe-6de633ed4826>

Villas-Boas A, ’China is Trying to Prevent Deepfakes With New Law Requiring That Videos Using AI Are Prominently Marked’ *Business Insider* (30 November 2019) <www.businessinsider.com/china-making-deepfakes-illegal-requiring-that-ai-videos-be-marked-2019-11?r=US&IR=T>

Vincent J, 'Tom Cruise deepfake creator says public shouldn't be worried about 'one-click fakes'' *The Verge* (5 March 2021) <www.theverge.com/2021/3/5/22314980/tom-cruise-deepfake-tiktok-videos-ai-impersonator-chris-ume-miles-fisher>

Vocal Synthesis, ‘Jay-Z covers “We Didn’t Start The Fire” by Billy Joel (Speech Synthesis)’ (*YouTube*, 25 April 2020) <https://youtu.be/iyemXtkB-xk>

Wang R, and others, ‘*FakeTagger*: Robust Safeguards against DeepFake Dissemination via Provenance Tracking’ (2021) available at: <https://arxiv.org/abs/2009.09869v2>

Welcome to Chechnya (2020)

Wertheim J, and Hickey B, ‘The Mother of All Deepfakes’ *Sports Illustrated* (12 May 2021) <www.si.com/media/2021/05/12/deepfake-pompom-mom-daily-cover>

Westerlund M, ’The Emergence of Deepfake Technology: A Review’ (2019) 9(11) Technology Innovation Management Review 40

White House, 'Executive Order on Preventing Online Censorship' (*Trump White House* 28 May 2020) <https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-preventing-online-censorship/>

Wiggers K, ‘Fewer than 30% of Business Have Plan to Combat Deepfakes, Survey Finds’ (*Venture Beat,* 24 May 2021) <https://venturebeat.com/2021/05/24/less-than-30-of-business-have-a-plan-to-combat-deepfakes-survey-finds/>

Wilkerson L, ’Still Water Run Deep(fakes): The Rising Concerns of “Deepfake” Technology and Its Influence on Democracy and the First Amendment’ (2021) 86(1) Missouri Law Review 407

WIPO, ‘WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI)’ (13 December 2019) WIPO/IP/AI/2/GE/20/1, available at <www.wipo.int/export/sites/www/about-ip/en/artificial_intelligence/call_for_comments/pdf/ind_lacasa.pdf>

WITNESS, 'Deepfakes: Prepare Now (Perspectives from Brazil)’ (*WITNESS Media Lab*)

Wojewidka J, ‘The deepfake threat to biometrics’ (2020) 2 Biometric Technology Today 5

Woods L, ‘Digital Freedom of expression in the EU’ in Douglas-Scott S, and Hatzis N, (eds) *Research Handbook on EU Law and Human Rights* (Edward Elgar Publishing 2017)

Wu F, Ma Y, and Zhang Z, '"I Found a More Attractive Deepfaked Self": The Self-Enhancement Effect in Deepfake Video Exposure' (2021) 24(3) Cyberpsychology, Behavior, and Social Networking 173

Yang C, and others, 'Preventing DeepFake Attacks on Speaker Authentication by Dynamic Lip Movement Analysis' (2021) 16 IEEE Transactions on Information Forensics and Security 1841

Yazdinejad A, and others, ’Making Sense of Blockchain for Deepfake Technology’ in *2020 IEEE Globecom Workshops* (IEEE 2020) 1

Yildirim G, Seward C, and Bergmann U, ‘Disentangling Multiple Conditional Inputs in GANs’ (2019) ICCV 2019 Conference Workshop on Computer Vision for Fashion, Art and Design, <https://research.zalando.com/publication/generating_models_2019/>

Yin X, and Hassner T, 'Reverse engineering generative models from a single deepfake image' (*Facebook AI*, 16 June 2021) <https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image>

YouTube Help, ‘Submit a copyright takedown request’ (*YouTuber Help Centre*) <https://support.google.com/youtube/answer/2807622?hl=en-GB>

Zhao B, and others, ‘Deep Fake Geography? When Geospatial Data Encounter Artificial Intelligence’ (2021) 48(4) Cartography and Geographic Information Science 338

Zhao Y, and others, ‘Capturing the Persistence of Facial Expression Features for Deepfake Video Detection’ in Zhou J, and others (eds.) *Information and Communications Security 21st International Conference, ICICS 2019* (Springer 2019) 632

Websites

Binded <https://binded.com/>

Deliberate <www.deliberate.ai/>

FaceApp <www.faceapp.com >

Project Origin <www.originproject.info/>

Sentinel <https://thesentinel.ai/>

Thispersondoesnotexist <https://thispersondoesnotexist.com/>

Truepic <https://truepic.com/technology/>

Which Face is Real? <www.whichfaceisreal.com/index.php>