'Deepfakes as an Emerging Threat to Freedom and Democracy: Is the Law Prepared to Respond?'

DANIELLE S. VAN LIER
Dissertation submitted to The University of Edinburgh
LLM in Innovation, Technology and the Law
August 19, 2019

Abstract: Artificial intelligence has democratized the ability to create high-quality manipulated videos that were once limited to Hollywood special effects houses. The technology has potential to bring significant benefit to a number of fields, but it also poses significant risks. This paper examines the deepfakes phenomenon, the risks it poses, and how to combat those risks. It primarily address the questions from a United Kingdom perspective, focusing on England and Wales, while drawing on research and legal approaches from other jurisdictions, particularly the United States.

1. Introduction*

A pop star is drugged by her manager, falls into a coma, and is replaced by a hologram version of herself.[1] This is the premise of an episode of the dystopian show "Black Mirror" but is it so outside the realm of possibility? Movies have postulated similar concepts for years.[2] Modern digital technology, including deepfakes, is bringing these plots closer to reality.

Hollywood has long had technology to digitally insert people into motion pictures, to manipulate performances, and even to depict actors as nude when they did not perform as such. Recent advances in artificial intelligence ("AI") have enabled anyone with a reasonably powerful computer to create similar effects. Manipulated images — including sexual ones — have circulated online since the early days of the internet. But deepfakes are high-quality digital fakes that can be difficult to detect. The ease of creating them and the ability to spread them widely and rapidly increase their potential for harm.

---

* Please note that this is the version submitted for purposes of my dissertation and, as such, contains a number of typos and errors common in student papers. Due to timing, workload, and the university's assessment regulations, I have not yet had a chance to produce a cleaner version.

Many thanks to my SAG-AFTRA colleagues who helped spot my typos: Zino Macaluso, National Director/Senior Counsel, Professional Representatives; Sarah Luppen Fowler, Associate General Counsel; Talin Galoosian, Counsel; Matt Blackett, Counsel

[1] Adi Robertson 'Rachel, Jack and Ashley Too is Black Mirror's stab at a feel-good teen comedy' (The Verge Jun 7, 2019) <www.theverge.com/2019/6/7/18653290/black-mirror-review-season-5-rachel-jack-ashley-netflix-charlie-brooker-miley-cyrus> accessed July 16, 2019

[2] IMDb, 'S1m0ne' <www.imdb.com/title/tt0258153/> accessed 18 Aug 2019 (a producer replaces his lead actress with a digitally-created actress); IMDb, 'The Congress' <www.imdb.com/title/tt1821641/> accessed 18 Aug 2019 (an actress licenses away her likeness when she becomes tired of acting)

© 2019 Danielle S. Van Lier

As deepfake technology spreads and becomes more accessible, the dangers it poses are becoming more apparent. These range from the intensely personal to the global. They expose individuals to potential exploitation, blackmail, defamation, and other harms. On a societal level, they represent a new and pernicious form of propaganda that can cause us to question even truthful information. But the technology behind deepfakes is not inherently evil – it can be used for numerous beneficial purposes.

Lawmakers are wrestling with how to address deepfakes. In November 2018, the Law Commission for England and Wales published a Scoping Report on Abusive and Offensive Online Communications ("Scoping Report") following a "six-month project to analyze the current criminal law relating to abusive and offensive online communication".[3] The Commission's review included identifying "the impacts of abusive and offensive online behavior," including deepfakes[4] It included meetings with stakeholders, such as abuse victims and organizations that assist them.[5] The Scoping Report's findings are consistent with the scholarship around online abuse, as discussed throughout this paper.

This paper will examine the deepfakes phenomenon, the risks it poses, and what can be done to combat those risks. It will primarily address the legal issues from a United Kingdom ("UK") perspective, focusing on England and Wales, while drawing on research and legal approaches from other jurisdictions, particularly the the United States ("US") where much of the deepfakes-related scholarship and reporting originates.

Section 2 answers the threshold question, "what are deepfakes, anyway?" It looks at the technology/ies used to create deepfakes as well as the audio and video output embodied within the term "deepfakes." Section 3 discusses disinformation, generally, how it is spread, and how deepfakes exacerbate the problem. Section 4 looks at the use of deepfakes to exploit individuals, particularly with non-consensual sexual content. Section 5 addresses deepfakes as a threat to democratic freedoms, including their potential impact on national security, public safety, and journalism. Sections 4 and 5 address necessary changes in the law; section 6 offers additional legal, policy, and technical solutions.

2. What are "deepfakes," anyway?

A. Defining "deepfakes"

A Reddit user, "@deepfakes," posted one of the first pornographic deepfakes — a video of actress Gal Gadot's face on the body of a pornography actress.[6] "A second Reddit user,

[3] Law Commission *Abusive and Offensive Online Communications: A Scoping Report* (Law Com No 318, 2018) para 1.1

[4] ibid para 3.2

[5] ibid para 3.4

[6] Samantha Cole, 'AI-Assisted Fake Porn Is Here and We're All Fucked' (*Vice*, 17 Dec 2017) <www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn> accessed 31 Jul 2019

© 2019 Danielle S. Van Lier

@deepfakeapp, created the first widely distributed deepfake software, called "FakeApp."[7] FakeApp built on Google's TensorFlow tool.[8] Deepfake" (or "deep fake"), a portmanteau of "deep learning" and "fake," is now the name for a broad category of content.[9]

All deepfakes share certain key characteristics: the use of AI (usually a deep learning algorithm), automation, and falsification and/or the potential to deceive.[10] This formulation — visually, a Venn diagram where the overlap represents deepfakes — would encompass most deepfakes while eliminating manually altered visual content.[11] The latter can be just as damaging, but distinguishing them ensures precise definitions.[12]

"Deepfake" is used broadly in this paper: i) as an adjective to describe: (a) the AI-driven applications that create altered audio and video clips and (b) the output from those applications, and ii) as a noun, to refer to the AI-generated content. The latter includes content in which one or more person(s) are digitally inserted, altered, or manipulated to say or do something they did not actually say or do in the manner depicted.

B. Image alteration technology

Throughout history, people have doctored images for purposes ranging from benign to satirical to malicious. For example: a famed portrait of US President Abraham Lincoln used a body model;[13] a 1930s photograph of Joseph Stalin was doctored to remove someone he disfavored;[14] and in 2014, a farmer in China attempted to blackmail politicians by digitally

---

[7] Dave Gershgorn 'Google Gave The World Powerful AI Tools, And The World Made Porn With Them' (*Quartz*, 7 Feb 2018) <https://qz.com/1199850/google-gave-the-world-powerful-open-source-ai-tools-and-the-world-made-porn-with-them/> accessed 17 Aug 2019

[8] ibid

[9] Hazel Baker 'Making a 'deepfake': How creating our own synthetic video helped us learn to spot one' (*Press Gazette*, 11 Mar 2019) <www.pressgazette.co.uk/making-a-deepfake-how-creating-our-own-synthetic-video-helped-us-learn-to-spot-one/> accessed 18 Aug 2019; James Vincent, 'Why we need a better definition of "deepfake"' (*The Verge*, 22 May 2018) <www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news> accessed 18 Aug 2019

[10] Vincent (n9)

[11] ibid

[12] ibid

[13] Fourandsix Technologies, Inc. 'Circa 1860' (*Photo Tampering Throughout History*) <http://pth.izitru.com/1860_13_00.html> accessed 18 Aug 2019

[14] Fourandsix Technologies, Inc. 'Circa 1930' (*Photo Tampering Throughout History*) http://pth.izitru.com/1930_13_00.html

© 2019 Danielle S. Van Lier

inserting them into pornographic photographs.[15] Today, consumer-level digital imagery and editing tools make it easy for anyone with a modicum of skill to convincingly edit photographs.[16]

Since the early days of filmmaking, technology has allowed motion picture studios to add and modify people in films. These technologies constantly evolve. 1961's "The Parent Trap" used a simple doubling technique to depict Hayley Mills as two different characters in the same shot.[17] 1994's "Forrest Gump" featured Tom Hanks in historical footage of major world leaders. Filmmakers used voice doubles and digitally manipulate the leaders' mouths to match the dialogue.[18] 2006's "Superman Returns" used old footage and audio to depict Marlon Brando as Superman's father.[19] Brando's appearance was brief but laid the groundwork for future digital resurrections. More recently, "Rogue One: A Star Wars Story" digitally resurrected Peter Cushing to portray a lead role.[20] An actor with a similar build, stature, and manner of speaking wore motion capture equipment during production and was digitally replaced by Cushing in post-production.[21]

Deepfake technology is an evolution of these production technologies. Open-source AI tools, such as TensorFlow, give the public capabilities once reserved to the film and video game industries.[22] With a reasonably powerful personal computer anyone can do what once cost millions of dollars and took hundreds of hours. The quality is not as high, but deepfakes are of a high enough quality to be difficult to detect.[23]

---

[15] Fourandsix Technologies, Inc. 'June 2014' (*Photo Tampering Throughout History*) http://pth.izitru.com/2014_06_00.html

[16] Robert Chesney and Danielle Citron, 'Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics' (*Foreign Affairs*, Jan-Feb 2019) <www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war> accessed 18 Aug 2019 (also available Academic OneFile, http://link.galegroup.com/apps/doc/A566263296/AONE?u=ed_itw&sid=AONE&xid=9a729105))

[17] Allie Townsend 'The Parent Trap — Doubling', A Brief History of Movie Special Effects Time <http://content.time.com/time/photogallery/0,29307,2055255_2247328,00.html> visited August 4, 2019

[18] Bart Mills 'In "Forrest Gump," Historical Figures Speak For Themselves' *Chicago Tribune* (Chicago, 8 July 1994) <www.chicagotribune.com/news/ct-xpm-1994-07-08-9407080087-story.html> visited August 4, 2019

[19] Michael Kunkes, 'Brando Talks! and Superman Listens' (CineMontage, 1 July 2006) <https://cinemontage.org/brando-talks-and-superman-listens/> visited August 4, 2019

[20] Dave Itzkoff, 'How "Rogue One" Brought Back Familiar Faces' *The New York Times* (New York, 27 Dec 2019) <www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html> accessed 18 Aug 2019

[21] ibid; Kevin Lincoln, 'How did Rogue One Legally Re-create the Late Peter Cushing?' (*Vulture* 16 Dec 2016) <www.vulture.com/2016/12/rogue-one-peter-cushing-digital-likeness.html> accessed 18 Aug 2019

[22] Douglas Harris, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You' (2019) 17 Duke Law & Technology Review 99

[23] Chesney and Citron, 'Deepfakes and the New Disinformation War' (n 16)

© 2019 Danielle S. Van Lier

C. How deepfakes are created

Understanding how deepfakes are created and the "artifacts" the technology leaves behind is important to combatting them. It is also critical to recognize how quickly the technology is evolving, making it difficult to employ technological measures against deepfakes them.

i. Deepfakes and machine learning

AI is "the study of agents that perceive the world around them, form plans, and make decisions to achieve their goals."[24] Machine learning ("ML") is a subfield of AI that encompasses "computer processes capable of learning from data to make ever-more accurate decisions and predictions."[25] It enables a computer to learn on its own – to "identify patterns in observed data, build models that explain the world, and predict things without having explicit pre-programmed rules and models."[26] Deep learning is a recent, more specialized form of ML that uses "neural nets" — a form of processing designed to function like neurons in the human brain.[27] Deep-learning AIs learn to detect abstract concepts without pre-set data.[28] They learn through iteration, starting out wildly inaccurate until they fine-tune and master the task over time.[29]

Generative adversarial networks ("GAN") are the deep-learning tools used to create deepfakes. GANs pit two competing neural networks against each other to learn and improve.[30] One network generates samples and the other uses training data and the generator's samples to predict between two different outcomes, such as "real" or "fake."[31] The "generator" and "discriminator" gradually train each other with the former generating increasingly accurate samples (or increasingly realistic images) in an attempt to beat the latter.[32] GAN's are able to "determine, refine, and innovate criteria, patterns, and strategies by themselves" — they can

---

[24] Vishal Maini, 'Machine Learning for Humans' (*Medium*, 19 Aug 2017) <https://medium.com/machine-learning-for-humans/why-machine-learning-matters-6164faf1df12> accessed 16 Aug 2019

[25] John Fletcher, 'Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance' (2018) 70 Theatre Journal 455, 457-58

[26] Maini (n 24)

[27] Fletcher (n 25) p 458

[28] Fletcher (n 25) p 458-9, Maini (n 24)

[29] Fletcher (n 25) p 459

[30] Fletcher (n 25) p 459

[31] Fletcher (n 25) p 459, Rani Horev. 'Style-based GANs – Generating and Tuning Realistic Artificial Faces' (*LyrnAI*, 26 Dec 2018) <www.lyrn.ai/2018/12/26/a-style-based-generator-architecture-for-generative-adversarial-networks/> accessed 23 Jul 2019

[32] Fletcher (n 25) p 459, Horev (n 31)

© 2019 Danielle S. Van Lier

learn and develop without human intervention.[33] This extraordinary power has sparked concerns among technologists and ethicists.[34]

To create a deepfake, a GAN is trained with a large data set, including images of the subject person and the target video, which it uses to generate a new video.[35] Deepfakes can be created in a matter of hours using an adequately powerful computer.[36]

D. Types of Deepfakes

Deepfakes fall into three primary categories. The most prevalent, especially in pornography, remains the "face swap" deepfake, like the Gal Gadot video mentioned above.[37] "Puppet master" deepfakes use the movements, such as head and eye movements, and facial expressions, of an unseen actor or "puppet master" to animate the target. [38] "Lip-sync" deepfakes reanimate the target's mouth to sync up to a new audio track, often spoken by an impersonator or generated by a computer.[39]

The same technologies can be used to create fake audio. Manually creating audio fakes is time-consuming and requires a large database of sound fragments that can be rearranged and combined to simulate speech. AI allows the creation of increasingly convincing voice facsimiles with increasingly smaller samples. These technologies can use short samples of an individual's speech to generate audio of them speaking completely new sentences.[40]

---

[33] Fletcher (n 25) p 459

[34] Fletcher (n 25) p 459

[35] Siwei Lyu, 'This ingenious algorithm outsmarts deepfakes' (*Fast Company*, 27 Jun 2019) <www.fastcompany.com/90370113/this-ingenious-algorithm-outsmarts-deepfakes> accessed 7 Aug 2019

[36] Cole, 'AI-Assisted Fake Porn' (n 6)

[37] Shruti Agarwal and others, 'Protecting World Leaders Against Deep Fakes' (CVPR Workshops, Long Beach, 2019) <http://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.html> accessed 18 Aug 2019. See also Drew Harwell, 'Top AI researchers race to detect 'deepfake' videos: 'We are outgunned'' (*The Washington Post*, 12 June 2019) <www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/> accessed13 Aug 2019

[38] Harwell (n 37)

[39] Agarwal (n 37)

[40] Sebastian Anthony, 'Adobe demos "photoshop for audio," lets you edit speech as easily as text' (*Ars Technica*, 7 Nov 2016) <https://arstechnica.com/information-technology/2016/11/adobe-voco-photoshop-for-audio-speech-editing/> accessed 19 Aug 2019; Bahar Gholipour, 'New AI Tech Can Mimic Any Voice' (*Scientific American*, 2 May 2017) <https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/> accessed 19 Aug 2019; Sercan Arik and others, "Neural Voice Cloning with a Few Samples." (NeurIPS Conference, Montréal, Dec 2018). <http://papers.nips.cc/paper/8206-neural-voice-cloning-with-a-few-samples> accessed 18 Aug 2019 See also Lyrebird 'Vocal Avatar' <https://lyrebird.ai/vocal-avatar>

© 2019 Danielle S. Van Lier

Researchers are introducing new, innovative, and sometimes abusive, technologies at a rapid pace. For example the following technologies were released since early 2018:

- Full-body deepfakes technology that maps and swaps the subject's entire body, including their clothes, onto the target person. [41]
- A "do as I do" motion transfer method that uses movement from a source in one video to animate a target in a second video.[42]
- "Deep video portraits" — a method to transfer the 3D head position and rotation and facial expression and movement from a source to a video portrait (*i.e.* the head and upper body).[43] In testing, roughly half of viewers thought the videos were real.[44]
- "Talking head models" created from a single photograph.[45] The technique can even animate paintings or other still images.[46]
- Facial animation using a still image and an audio clip with "(a) lip movements that are in sync with the audio and (b) natural facial expressions."[47]
- "DeepNude" — a now-defunct application that virtually removed clothing from women in photographs.[48]

---

accessed 18 Aug 2019 (featuring proof-of-concept audio clips of US presidents Barack Obama and Donald Trump)

[41] Dan Robitzski, 'These Full-Body Deepfakes are Like Nothing We've Ever Seen' (*Futurism,* 12 Dec 2018) <https://futurism.com/full-body-deepfakes> accessed 18 Aug 2019

[42] Caroline Chan and others 'Everybody Dance Now' (Unpublished Paper, ArXiv 2018) p 1 <https://arxiv.org/abs/1808.07371> accessed 18 Aug 2019

[43] Hyeongwoo Kim and others, 'Deep video portraits' (2018) 37 ACM Transactions on Graphics, (4), p 1 <https://doi.org/10.1145/3197517.3201283> (also available with videos at https://web.stanford.edu/~zollhoef/papers/SG2018_DeepVideo/page.html) accessed 18 Aug 2019

[44] ibid p 12

[45] Egor Zakharov and others, 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models." (Unpublished paper, ArXiv 2019) p 2 <https://arxiv.org/abs/1905.08233> accessed 18 Aug 2019

[46] ibid p 1

[47] Konstantinos Vougioukas, Stavros Petridis and Maja Pantic 'Realistic Speech-Driven Facial Animation with GANs.' (Unpublished paper, ArXiv 2019) p 1 <https://arxiv.org/abs/1906.06337> accessed 18 Aug 2019

[48] Samantha Cole, 'This Horrifying App Undresses a Photo of Any Woman With a Single Click' (*Vice*, 26 Jun 2019) <www.vice.com/en_us/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman> accessed 24 July 2019; Sigal Samuel, 'A guy made a deepfake app to turn photos of women into nudes. It didn't go well.' (*Vox*, 27 Jun 2019) <www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn> accessed 27 Jul 2019

© 2019 Danielle S. Van Lier

- StyleGAN, an application that generates artificial faces that looks realistic.[49] The application can generate other objects — such as cars, rooms, and cats — depending on the source data.[50]

E. Beneficial Uses for Deepfake Technology

Deepfakes have significant potential for abuse, but the technologies also have many potential positive applications. The most obvious application is in the entertainment industry where they can create high-quality digital reproductions and animation quicker and at a lower cost than current techniques. To illustrate, the deep video portraits technology has utility in the context of dubbing foreign language films, among other applications.[51] And the full-body deepfake technology can augment or replace costly motion capture in video games and animated movies.[52]

Several deepfake technologies have educational applications. The talking head technology can be used to create realistic avatars in virtual and augmented reality applications, which its creators believe "will democratize education, and improve the quality of life for people with disabilities."[53] The facial animation application can bring historic images to "life" to tell their stories. For example, the USC Shoah Foundation scans Holocaust survivors to create interactive "holograms" of them to teach about the Holocaust.[54] Deepfake technology can be used to create similar learning tools without the need for the individual's participation, something that may be useful in telling the stories of many historical figures.

Voice reproduction technologies have applications in education, customer service, and to aid the disabled. An Edinburgh company used AI to recreate the speech US President John F. Kennedy was planning to deliver on the day he was assassinated.[55] The company used text

---

[49] Jackson Ryan, 'This website uses AI to generate startling fake human faces' (*CNet,* 14 Feb 2019) <www.cnet.com/news/this-website-uses-ai-to-generate-startling-fake-human-faces/> accessed 23 July 2019; Tero Karras, Samuli Laine and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks." (Unpublished paper, ArXiv 2019) <https://arxiv.org/abs/1812.04948> accessed 18 Aug 2019 (For examples, see https://thispersondoesnotexist.com/)

[50] Karras (n 50) p 9

[51] Kim (n 44) p 2

[52] Robitzski (n 42)

[53] Egor Zakharov, 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models' (*YouTube* 21 May 2019) <https://youtu.be/p1b5aiTrGzY> accessed 5 Aug 2019

[54] Ellen Braunstein, 'At This Holocaust Museum, You Can Speak With Holograms Of Survivors' (*Jewish Telegraphic Agency*, 22 Jan 2018) <www.jta.org/2018/01/22/united-states/at-this-holocaust-museum-you-can-speak-with-holograms-of-survivors> accessed 6 Aug 2019

[55] BBC News 'John F Kennedy's Lost Speech Brought to Life' (16 Mar 2018) <www.bbc.com/news/uk-scotland-edinburgh-east-fife-43429554> accessed 28 July 2019; 'Artificial Intelligence Used to Recreate JFK's Dallas Speech That He Never Gave' *The Washington Post* (Washington DC 13 Apr 2018) www.washingtonpost.com/video/business/technology/artificial-intelligence-used-to-recreate-jfks-dallas-

© 2019 Danielle S. Van Lier

from Kennedy's past speeches and 116,000 snippets of speech from prior recordings.[56] The company is researching the use of AI for speech synthesis, including to allow those who lose their voices to communicate in their own voice.[57]

Deepfakes have tremendous beneficial applications; they become dangerous when they feed the natural human tendency to believe, remember, and share negative and novel information, including false information.[58] Humans are "biologically programmed to be attentive to things that stimulate: content that is gross, violent, or sexual, and gossip, which is humiliating, embarrassing, or offensive" and gravitate toward content that creates an emotional response.[59] The remainder of this paper will focus on the dark side of deepfakes and the threats they pose.

3. Deepfakes as tools of deception and disinformation

A. Fake news and how it spreads

"Fake news" or disinformation — "false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" [60] — has a long history, particularly in the political process.[61] Historians trace some stories as far back as the 12th century with fake news becoming more common as the printing press allowed widespread

speech-that-he-never-gave/2018/05/10/caad5f8a-3f33-11e8-955b-7d2e19b79966_video.html?utm_term=.4f13230fd713 accessed 28 July 2019

[56] BBC (n55); History News Network 'Artificial Intelligence Has Been Used To Recreate JFK's Dallas Speech That He Never Gave' (10 May 2018) <https://historynewsnetwork.org/article/169020> accessed 28 July 2019

[57] CereProc, 'CereVoice Me Voice Cloning Service' <https://www.cereproc.com/en/products/cerevoiceme> visited 28 July 2019

[58] Robert Chesney and Danielle Keats Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2018) 107 California Law Review (2019, forthcoming) <http://dx.doi.org/10.2139/ssrn.3213954>p 12

[59] danah boyd, 'Streams of Content, Limited Attention: The Flow of Information Through Social Media' (*Web2.0 Expo,* Nov. 17, 2009) < http://www.danah.org/papers/talks/Web2Expo.html> accessed 16 Aug 2019

[60] European Commission, 'A Multi-Dimensional Approach To Disinformation: Report Of The Independent High Level Group On Fake News And Online Disinformation' March 2018 p 10 <https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation>

[61] Jacob Soll, 'The Long and Brutal History of Fake News' (*Politico Magazine*,18 December 2016) <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535> accessed 19 Aug 2019. *See also* David Uberti 'The real history of fake news (*Columbia Journalism Review*, 15 Dec 2016) <https://www.cjr.org/special_report/fake_news_history.php> accessed 18 Aug 2019

© 2019 Danielle S. Van Lier

dissemination.[62] In the 17th century, Charles II even issued a proclamation prohibiting the writing, speaking, or publication of false news.[63]

The internet has further democratized information dissemination with the effect of removing major news outlets, which typically adhere to ethical standards in their reporting, as gatekeepers. Individuals can create and spread disinformation with internet access and a minimal degree of technical savvy. Social media aids it spread because it can "efficiently deliver messages to large groups of targeted people, much more effectively than" traditional communication methods, making it fertile ground for the spread of disinformation.[64]

People are bad at detecting fake stories, even when they are purely text-based. [65] Deepfakes likely will exacerbate this phenomenon because of the persuasive power of audio and video, which allow people to be firsthand witnesses to an event without the need to trust someone else's account.[66] People are predisposed to trust what we see and hear and are even less inclined to seek verification of video and audio accounts of events. [67]

B. The impact of disinformation

Concern over disinformation is increasing. In the UK, 84% of respondents believe fake news is a problem domestically and 78% believe it is a threat to democracy.[68] Additionally, 84% of UK respondents believe they come across fake news at least several times a month, although 79% felt somewhat or very confident they could identify it.[69] UK respondents feel the primary

[62] Soll (n 61)

[63] England and Wales. Sovereign (1660-1685 : Charles II). (1674). By the King. A proclamation to restrain the spreading of false news, and licentious talking of matters of state and government available at http://ota.ox.ac.uk/tcp/headers/B02/B02127.html. See also Kenan Malik, 'Fake News Has A Long History. Beware The State Being Keeper Of "The Truth"' *The Guardian* (London 10 February 2018) https://www.theguardian.com/commentisfree/2018/feb/11/fake-news-long-history-beware-state-involvement accessed 8 August 2019;

[64] Kurt Wagner, 'Facebook And Twitter Worked Just As Advertised For Russia's Troll Army' (*Recode*, 17 February 2018) <www.recode.net/2018/2/17/17023292/facebook-twitter-russia-donald-trump-us-election-explained> accessed 19 Aug 2019

[65] Holly Kathleen Hall, 'Deepfake Videos: When Seeing Isn't Believing' (2018) 27 The Catholic University Journal of Law & Technology 51, 56

[66] Chesney and Citron, 'Deepfakes and the New Disinformation War (n 16)

[67] ibid

[68] European Commission, 'Fake News And Disinformation Online: United Kingdom' (2018) <https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/flash/surveyky/2183> accessed 10 August 2019

[69] ibid

© 2019 Danielle S. Van Lier

responsibility to stop its spread lies with press and broadcasting management, journalists, and online social networks.[70]

Another poll found that 75% of adults in Great Britain believe fake news should be criminalized.[71] This result was split fairly evenly between those who "tend to agree" and those who "strongly agree," with only three percent strongly disagreeing.[72] Nearly 70% were worried fake news could influence the result of an election or referendum.[73]

Among social media users, 62% trust the news they get from social media at least a "fair amount."

C. The 'liar's dividend'

As deepfakes become more prevalent, they will prime the public to discount or even disregard video or audio evidence, to not believe what they see or hear.[74] This will create a "liar's dividend" where public figures, such as politicians, will find it easier to convince the public that real audio or video is false.[75] The public, likewise, will become less trusting of the news.[76] "If people can no longer believe what they see and what they hear, it's easy for political leaders to dismiss evidence-based negative coverage as" false.[77] In the US, President Donald Trump frequently denies comments that have been caught on video, such as the Access Hollywood video of him boasting about assaulting women.[78]

---

[70] ibid

[71] Ipsos MORI, 'Three-quarters would make spreading fake news a crime' (17 Apr 2019) <https://www.ipsos.com/ipsos-mori/en-uk/three-quarters-would-make-spreading-fake-news-crime> accessed 17 Aug 2019

[72] ibid

[73] ibid

[74] Chesney and Citron, 'Deepfakes and the New Disinformation War' (n 16)

[75] ibid

[76] ibid

[77] Rubina Madan Fillion, 'Fighting the reality of deepfakes' (Nieman Journalism Lab 2019) <https://www.niemanlab.org/2018/12/fighting-the-reality-of-deepfakes/> accessed 8 Aug 2019

[78] Maggie Haberman and Jonathan Martin 'Trump Once Said the "Access Hollywood" Tape Was Real. Now He's Not Sure' *The New York Times* (New York 28 Nov 2017) <https://www.nytimes.com/2017/11/28/us/politics/trump-access-hollywood-tape.html> accessed 15 Aug 2019. See also David Gilbert 'Trump Claims The Lester Holt Interview Where He Basically Admitted To Obstruction Is Somehow Fake' (*Vice*, 30 Aug 2018) <https://www.vice.com/en_us/article/ne5x3d/trump-lester-holt-james-comey-nbc> accessed Aug 15 2019; Chris Morran 'Trump Claims He Never Called Meghan Markle 'Nasty,' Demands Apology From CNN, Even Though Video Exists' (*Newsweek*, 2 Jun 2019) <https://www.newsweek.com/trump-claims-he-never-called-meghan-markle-nasty-demands-apology-cnn-even-1441478> accessed 15 Aug 2019

© 2019 Danielle S. Van Lier

People already question the veracity of the legitimate news they receive, including from mainstream outlets. In the aforementioned poll, 35% of British respondents believe they "would never get fake news on the BBC" with a similar number holding the opposite view.[79] Troublingly, nearly half of respondents in Scotland believe it is possible they would get fake news on the BBC with just 24% believing they never would.[80]

4. Deepfakes as a form of Image-Based Abuse

In the movie, "Nymphomaniac," filmmakers used digital technology to insert the film's stars into explicit sex scenes. The scenes were performed by body doubles who actually engaged in the sex acts and then were digitally replaced by the stars.[81] This was done with the actors' consent, pursuant to negotiated contracts. Deepfakes put this capability in anyone's hands and allows the creation of similar content without the subject's participation, consent, or even knowledge.

A. Sexually-explicit deepfakes and gender-based sexual violence

The most widespread deepfakes to-date are non-consensual sexually-explicit videos of women.[82] The videos are not limited to celebrities — people are creating videos of ex-intimates, co-workers, and others.[83] As one writer observed: "Google gave the world powerful AI tools, and the world made porn with them."[84]

Sexually-exploitative deepfakes fall on a continuum of image-based sexual abuse ("IBSA").[85] This continuum includes abusive behaviors like revenge pornography, sexualized extortion, and voyeurism.[86] All forms of IBSA "reduce victims to sexual objects that can be exploited and exposed;" they deny the victim agency over their intimate lives.[87] The harms IBSA victims suffer

---

[79] Ipsos MORI (n 71)

[80] ibid

[81] Scott Roxborough 'Cannes: "Nymphomaniac" Producer Reveals Graphics Are Used in 'Groundbreaking' Sex Scenes' (*The Hollywood Reporter*, 20 May 2013) <https://www.hollywoodreporter.com/news/cannes-nymphomaniac-producer-sex-scenes-525666> accessed 15 Aug 2019

[82] Russell Brandom, 'Deepfake Propaganda is Not a Real Problem' (The Verge 5 Mar 2019) <https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax> accessed 18 Aug 2019

[83] Danielle Keats Citron, 'Sexual Privacy' (2019) 128 Yale Law Journal 1870, 1922

[84] Gershgorn (n 7)

[85] Clare McGlynn, Erika Rackley and Ruth Houghton 'Beyond "Revenge Porn": The Continuum of Image-Based Sexual Abuse (2017) 25 Feminist Legal Studies 25, 26-29

[86] ibid p 28

[87] Citron (n 83) p 1924

© 2019 Danielle S. Van Lier

—psychological harms, professional harm, harms to personal relationships and harms to reputation — are similar to other forms of sexual abuse.[88]

Sexually-exploitative deepfakes bear similarities to "revenge pornography," providing a useful reference point for analysis. However unlike revenge pornography, deepfakes do not depict the person's actual body and thereby create "a sexual identity not of the individual's making."[89] Deepfakes are dehumanizing because "a single aspect of one's self eclips[es] one's personhood."[90] They exploit the victim's sexual identity for other's gratification and force victims, particularly women, into virtual sex.[91] They even can turn rape threats into graphic videos depicting the act.[92]

Victims can suffer serious and sustained mental health effects.[93] This includes anxiety, panic attacks, anorexia, and depression.[94] Victims can suffer from low self-esteem and feelings of worthlessness.[95] And these feelings can grow over time, even leading to suicide.[96]

Some victims experience visceral fear and may simply withdraw from public life, particularly online activities.[97] Some do not feel safe leaving their homes.[98] The fear and risk is exacerbated when an individual is doxxed — when personal information such as contact information is released — along with the video posting.[99] Doxxing also can lead to ongoing harassment.[100]

IBSA is stigmatizing and victims may suffer professional and reputational consequences. They may lose jobs or have difficulty finding future employment.[101] In today's job market, a

---

[88] ibid

[89] Citron (n 83) p 1921

[90] Citron (n 83) 1925

[91] Chesney Citron p 17

[92] Chesney Citron 18, Citron (n 83) p 1924

[93] Clare McGlynn and Erika Rackley 'Image-Based Sexual Abuse' (2017) 37 Oxford Journal of Legal Studies 534

[94] Danielle Keats Citron and Mary Anne Franks, 'Criminalizing Revenge Porn' (2014) 49 Wake Forest Law Review 345, 345-391 p 351; David Ryan, 'European Remedial Coherence In the Regulationo Of Non-Consensual Disclosures of Sexual Images' (2018) 34 Computer Law & Security Review 1053, 1055; Citron (n 83) p 1926

[95] Ryan (n94)

[96] Citron and Franks (n 94) p 351

[97] ibid p 352; Citron (n 83) p1925

[98] Citron and Franks (n 94) p 351

[99] McGlynn and Rackley (n 93) p 545

[100] ibid

[101] Ryan (n 94); McGlynn and Rackley (n 93) p 545

© 2019 Danielle S. Van Lier

candidate's online reputation is crucial. Studies have showed that nearly 80% of employers do online research about candidates and 70% rejected candidates based on their findings.[102] Search results that include deepfakes may cost a victim interviews.[103] Victims may lose business opportunities, friendships and even romantic opportunities.[104] This also is true for actresses whose job opportunities and reputations can be harmed when they are involuntarily depicted in nude or sexual scenes, harming the image they seek to portray.[105]

Troublingly, some authorities dismiss the harm caused by manipulated images, as not having "the potential to cause the same degree of harm as the disclosure of images that record real private, sexual events."[106] Reports from victims and studies contradict this.[107]

B. Blackmail and harassment

The potential for reputational harm make deepfakes a likely tool for blackmail and extortion. Negative information spreads rapidly online, making the risk and the fear of reputational damage high, even if the video can eventually be debunked.[108]

Deepfakes also increase the risk of so-called sextortion, a form of blackmail or extortion involving threats "to release sexually explicit images of the victim if the victim does not engage in further sexual activity."[109] Typically, the perpetrator seeks additional nude photographs and also threatens further harm if the sextortion is disclosed, coercing the victim's silence.[110] The victims are nearly always women and many are underage.[111] Some operators of websites hosting non-consensual sexual content attempt sextortion when victims seek to have the content removed.[112] Deepfakes add a new dimension because the perpetrator – who is almost always a male – can falsify sexually explicit videos to be used in the sextortion attempt.

---

[102] Citron and Franks (n 94) p 352

[103] ibid p 352; Citron (n 83) 1928

[104] Chesney and Citron "Looming Challenge" (n 58) p 18

[105] Dave McNary 'SAG-AFTRA, Adam Schiff Express Alarm on "Deep Fake" Technologies' (*Variety,* 6 May 2019) <https://variety.com/2019/digital/news/sag-aftra-adam-schiff-alarm-deep-fake-technologies-1203206561/> accessed 7 Aug 2019

[106] McGlynn, Rackley and Houghton (n 85) p 34 citing (ministry of justice letter)]

[107] ibid p 34; Law Commission (n 3)

[108] Robert Chesney and Danielle Citron 'Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?' (Lawfare 21 Feb 2018) <https://www.lawfareblog.com/deep-fakes-looming-crisis-national-security-democracy-and-privacy> accessed 4 Aug 2019

[109] Citron (n 83) p 1915

[110] Citron (n 83) p 1916

[111] Citron (n 83) p 1916

[112] Citron (n 83) p 1924

© 2019 Danielle S. Van Lier

C. Civil Law Remedies for Deepfake Victims

There remains little academic literature on whether and how the law does, can, and/or should address deepfakes, especially in the UK. There is deeper scholarship on other forms of IBSA, particularly revenge pornography, with much based in the US. Accordingly, this subsection will borrow from that scholarship.

The law has addressed IBSA on an ad hoc basis, to the extent it has addressed it at all.[113] There is some merit to an incremental approach when attitudes toward an issue are continuing to evolve.[114] As the technology has matured and the harms of IBSA have become apparent, a comprehensive solution is necessary.[115]

i. Invasion of Privacy

Individuals have a fundamental right "to respect for. . . private and family life."[116] There is little question that deepfakes invade individual's privacy from the victim's perspective; the legal status is less clear. A few articles have raised privacy as potential claim, quickly dismissing it because deepfakes do not depict the actual individual.[117] Generally, these articles have examined US law.

English law does not recognize the broad privacy torts recognized in the US.[118] In *Wainwright and another v. Home Office*, Lord Hoffmann expressly rejected a "previously unknown tort of invasion of privacy" similar to the US approach, explaining that prior privacy invasions had been remedied under other legal theories rather than a general "invasion of privacy" principle.[119] To the extent it exists, the breach of privacy tort stems from "the protection of human autonomy and dignity — the right to control the dissemination of information about one's private life and the right to the esteem and respect of other people."[120]

There may be some limited recourse in the EU General Data Protection Regulations, enshrined in UK law through the Data Protection Act 2018.[121] The "right to erasure" may allow a victim to

---

[113] Citron (n 83) p 1939, 1944

[114] ibid p 1944

[115] ibid p 1944

[116] Human Rights Act 1998 art 8

[117] Chesney and Citron "Looming Challenge (n 58) p 36; Citron (n 83) 1939; Russel Spivak '"Deepfakes": The Newest Way To Commit One Of The Oldest Crimes' (2019) 3 Georgetown Law Technology Review 339, 377-381

[118] *Wainwright and another v. Home Office* [2003] UKHL 53, [2004] 2 AC 406 at [18]

[119] *Wainright* [19] – [35]

[120] *Campbell v Mirror Group Newspapers Ltd* [2004] UKHL 22; [2004] 2 AC 457 at [51].

[121] Nicholas Schmidt, 'Privacy law and resolving 'deepfakes' online' (*Privacy Perspectives*, 30 Jan 2019) <https://iapp.org/news/a/privacy-law-and-resolving-deepfakes-online/> accessed 18 Aug 2019;

© 2019 Danielle S. Van Lier

request removal of the content or they may be able to object to the processing of their data.[122] Additionally, an argument could be made that deepfakes are "personal data" because they "relate to an identified or identifiable natural person," even if the depiction is false.[123] But these theories seem tenuous. It is more likely that deepfake victims will not find adequate recourse in the context of privacy torts and will need to resort to other legal theories.

ii. Defamation

The inherent falsity of deepfakes make defamation a viable approach. In the UK, defamation was historically a common law tort but it was reformed by the Defamation Act 2013.[124] It is a strict liability tort with no specific motive required.[125] A statement, which can include pictures or visual images, is defamatory if it "caused or is likely to cause serious harm to the reputation of the claimant."[126] There are four potential defenses, none of which are likely applicable to deepfakes.[127] Operators of websites that host deepfakes have defenses in addition to the those discussed below.[128] However, they are not absolute and are conditioned on the operator taking certain steps set forth by regulation.[129]

The easiest way to defeat a defamation claim is to add a clear disclaimer or watermark indicating the deepfake's falsity. A video so identified is less likely to harm a victim's reputation, a threshold element of the claim, even if it still causes other harms. Defamation is also limited by the single publication rule which provides that a cause of action accrues upon the first publication of the defamatory statement.[130] This means the statute of limitation may expire before a victim even discovers the deepfake.

iii. Copyright Infringement

Copyright infringement is an ineffective remedy in the fight against deepfakes. The Copyright, Designs and Patents Act 1988 "CDPA" grants owners certain exclusive rights in their works.[131]

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC ('GDPR') art 17

[122] ibid

[123] ibid; GDPR art 4(1)

[124] Law Commission (n 3) par 11.98

[125] ibid par 11.103

[126] Defamation Act 2013 ss 1(1), 15

[127] Defamation Act 2013 ss 2 – 4, 7

[128] Defamation Act 2013 s 5

[129] Defamation Act 2013 s 5; The Defamation (Operators of Websites) Regulations 2013 SI 2013 No 3018

[130] Defamation Act 2013 s 8(3)

[131] Copyright, Designs and Patents Act 1988 ("CDPA") s 16(1)

© 2019 Danielle S. Van Lier

To state a claim for copyright infringement, the copyright owner must prove someone, without authorization, did any of the acts restricted by the copyright. [132]

The primary impediment to a copyright infringement claim is ownership of the underlying work(s). The producer of the pornographic video might have recourse but that would not benefit the victim. A growing number of pornographic videos are being created and released specifically for the purpose of creating deepfakes, reducing the infringement risk. Even if the photograph(s) used are selfies or the victim otherwise owns the copyright, a successful infringement claim is unlikely as most deepfake technologies use hundreds of photos to synthesize the individual's likeness, making it difficult to argue that a copyright actually has been infringed.

Section 85 of the CDPA provides a limited right to privacy in certain photographs or films commissioned for "private and domestic purposes."[133] Specifically, the subject has a right not to prevent the photograph's release to the public.[134] Section 85 may be useful in the context of revenge pornography, but is likely inapplicable to deepfakes.

iv. Inadequacy of civil remedies

Civil remedies currently are inadequate to address deepfakes. At best, they might offer modest monetary relief. But litigation is resource-, time-, and cost-intensive, which might prevent some victims from pursuing civil remedies, even if they have strong cases.[135] Other victims, particularly high-profile ones, may be dissuaded if they are unable to proceed pseudonymously.[136] Publicity from litigation has the paradoxical effect of increasing interest and driving more traffic to the videos — the exact opposite of what the victim wants.[137]

Online anonymity and other efforts taken by perpetrators to hide their identities are another impediment to adequate recourse.[138] The defendant may difficult or even impossible to locate and may lack assets to offer the victim meaningful redress.[139] Even if the a victim can obtain monetary damages or take-down orders against the creator or host, it is unlikely to serve the primary goal of litigation — to remove the videos from the internet and to repair their sullied

[132] CDPA s 16(2)

[133] CDPA s 85

[134] ibid

[135] Citron (n 83) p 1930

[136] Citron (n 83) p 1930; McGlynn and Rackley (n 93) p 559

[137] Ryan (n 94) 1056

[138] Citron (n 83) p 1929

[139] Ryan (n 94) 1056; McGlynn and Rackley (n 93) p 559

© 2019 Danielle S. Van Lier

reputation.[140] The lack of meaningful consequences makes these laws an inadequate deterrent to the creation and dissemination of deepfakes.

D Potential Criminal Penalties for Deepfake Creators

Criminal law is better suited to addressing the creation and spread of deepfakes. Criminal law penalties — the threat of jail time and a criminal record — is more likely to act as a deterrent than monetary and injunctive relief. Additionally, law enforcement has more robust investigative powers than would be available to a civil plaintiff.[141] Unfortunately, deepfakes do not fall neatly into existing criminal law and presently are not criminalized under English Law.[142]

i. Harassment

Online behavior, such as posting deepfakes, may be actionable as harassment if it otherwise meets the legal elements.[143] The Protection from Harassment Act 1997 ("PFHA") provides for imprisonment of up to six months and a fine upon conviction.[144] The victim also can pursue a civil action seeking damages "for (among other things) any anxiety caused by the harassment and any financial loss resulting from the harassment" plus injunctive relief.[145] In 2018, a man was convicted of harassment involving fake photos of a female intern he posted online, putting her job at risk.[146] The judge acknowledged the victim's current and future harm and anxiety when he sentenced the man to only 16 weeks in jail and a fine. This prosecution further illustrates a weakness in current law — the victim potentially will suffer life-long harm but the perpetrator will face minimal consequences.

Additionally, harassment laws are of limited efficacy because they require a repeated "course of conduct."[147] Under the PFHA, this means "conduct on at least two occasions in relation to" the victim or, where there are multiple victims, "conduct on at least one occasion in relation to each" victim. A single posting can cause harm; even if it goes viral, the poster committed a single act that would not constitute harassment.[148]

---

[140] McGlynn and Rackley (n 93) 559

[141] Citron (n 83) p 1929

[142] McGlynn, Rackley and Houghton (n 85) p 34

[143] Law Commission (n 3) para 8.7

[144] Protection from Harassment Act 1997 s 2

[145] Protection from Harassment Act 1997 s 3

[146] Cheyenne Roundtree 'City worker, 25, who tried to 'destroy' 22-year-old intern who rejected his advances by posting fake porn pictures of her on the internet is jailed for four months' *Daily Mail* (London 1 May 2018) <https://www.dailymail.co.uk/news/article-5678707/City-worker-25-tried-destroy-intern-rejected-jailed-four-months.html> accessed 19 Aug 2019

[147] Citron and Franks (n 94) p 365

[148] Citron and Franks (n 94) p 366

© 2019 Danielle S. Van Lier

ii. Revenge Pornography

There are many similarities between deepfakes and revenge pornography, warranting examination of whether existing revenge pornography laws address deepfakes or can easily be amended to include them. Section 33 of the Criminal Justice and Courts Act 2015 ("CJCA") provides that the non-consensual disclosure of a "private sexual photograph or film" is a criminal offense if done with the intention of causing distress.[149] This includes both physical photographs and "data. . . which is capable of conversion into an image," such as a digital photograph or video file.[150]

The CJCA covers images that have been altered.[151] However both the CJCA and its prosecutorial guidelines exclude images that have "become private and sexual. . . as a result of the alteration or combination."[152] According to the guidelines the CJCA does not cover "transpos[ing] the head of a former partner onto a sexual photograph of another person," as in "sexualized photoshopping" or "[i]mages which are completely computer generated but made to look like a photograph or film," such as deepfakes.[153]

The Abusive Behaviour and Sexual Harm (Scotland) Act 2016 ("ABSHA"), while similar in many respects, is broader. It includes photographs or films that "appear[] to show" a person "in an intimate situation."[154] However, the definitions of "film" and "photograph" require the content be captured by making a recording or by photography.[155] The explanatory notes indicate that the ABSHA does not apply to wholly digital content, particularly "material that looks like a photograph or film but does not in fact contain any photographic element (for example, because it had been generated entirely by computer)," such as deepfakes. [156]

Even if these Acts covered deepfakes, they would still be inadequate. The penalty under the CJCA is only two years plus a fine, an improvement over prior limits of 6-months imprisonment and a £5000 fine.[157] This remains low given the potential harms victims experience. The

---

[149] Criminal Justice and Courts Act 2015 ('CJCA') s 33

[150] CJCA s 34(8)(b)

[151] CJCA s 34(5)

[152] CJCA s 35(5)(b), (c); Crown Prosecution Service, 'Revenge Pornography - Guidelines on prosecuting the offence of disclosing private sexual photographs and films' (24 Jan 2017) <https://www.cps.gov.uk/legal-guidance/revenge-pornography-guidelines-prosecuting-offence-disclosing-private-sexual> accessed 18 Aug 2019. See also J Beyens and E Lievens 'A legal perspective on the non-consensual dissemination of sexual images: Identifying strengths and weaknesses of legislation in the US, UK and Belgium' International Journal of Law, Crime and Justice 47 (2016) 31-43, at 37.

[153] Crown Prosecution Service (n 160)

[154] Abusive Behaviour and Sexual Harm (Scotland) Act 2016 ('ABSHA') s 2(1)

[155] ABSHA s 3(2)

[156] Explanatory Notes Abusive Behaviour and Sexual Harm (Scotland) Act 2016

[157] Ryan (n 94) p 1059; CJCA s 33(9)

© 2019 Danielle S. Van Lier

maximum sentence under the ABSHA is 5 years, which better serves the deterrent function.[158] The ABSHA also criminalizes the threat to disclose images, which is a positive feature.[159] Problematically, these Acts require that the perpetrator intend to cause distress, which diminishes both the victim's sexual agency and the severity of the offense.[160] The creator or distributor of a deepfake is more likely to intend to profit or to satisfy a prurient interest than to cause distress.

iii. Communications Act 2003

The Scoping Report reviewed whether the Malicious Communications Act 1988 ("MCA") or Section 127 of the Communications Act 2003 ("Section 127") might be applicable to online offensive content.[161] Like the CJCA, the intent requirement of the MCA would limit its applicability to deepfakes. Section 127, however, makes it a criminal offense to send or cause to be sent "by means of a public electronic communications network a message or other matter that is grossly offensive or of an indecent, obscene or menacing character."[162] It is also an offense to similarly send or cause to be sent a message "that he knows to be false."[163] It is an extremely broad offense covering a broad swath of communication methods.[164] If convicted, the defendant may be imprisoned for up to six months and/or fined.[165] On its face, Section 127 would seem applicable to most deepfakes, particularly when accompanied by threats, harassment or doxing. However, as with harassment, the penalties are too low for it to have significant deterrent effect.

E. The role of Internet platforms

The internet platforms that host or link to deepfakes are in the best position to stem their distribution and are better positioned to pay any monetary damages that may be awarded to a victim. But platforms have broad immunity under current law.

i. The platform safe harbors

The European Union's E-Commerce directive, transposed into UK law through the Electronic Commerce (EC Directive) Regulations 2002, provide safe harbors if a platform is a "mere

[158] Ryan (n 94) p 1061; ABSHA s 2

[159] ABHSA s2

[160] CJCA s 33(1) ; McGlynn, Rackley and Houghton (n 85) p 38

[161] Law Commission (n 3)

[162] Communications Act 2003 s 127(1)

[163] ibid s127(2)

[164] Law Commission (n 3) para 4.63

[165] Communications Act 2003 s 127(3)

© 2019 Danielle S. Van Lier

conduit" or simply provides hosting services.[166] A platform is not liable for content transmitted over its services if the platform "(a) did not initiate the transmission; (b) did not select the receiver of the transmission; and (c) did not select or modify the information".[167] A hosting service provider is not liable if the provider did not have actual or constructive knowledge of the unlawful content or conduct or, once it becomes aware, if "it acts expeditiously to remove or to disable access to the information."[168]

ii. Voluntary steps taken by the platforms

Many social media sites have voluntarily taken steps to stem the proliferation of pornographic deepfakes on their platforms. Several that were among the earliest and most prominent repositories for pornographic deepfakes quickly updated their user policies or used existing policies to remove the nonconsensual videos. Reddit, where deepfakes first surfaced, updated its rules against nonconsensual sexual imagery to include deepfakes."[169] Twitter and other social media platforms soon followed suit.[170] Even pornography site Pornhub added deepfakes to its prohibited content.[171]

Predominantly, platforms rely on user reports or keyword-tracking to detect and remove deepfakes.[172] GIF-hosting site Gfycat, a popular platforms for deepfakes, takes a more aggressive approach and uses AI to identify and flag newly-uploaded deepfakes.[173] One tool recognizes the individual depicted in the GIF.[174] The second tool searches the internet for the

---

[166] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ("E-commerce directive") art 12

[167] Electronic Commerce (EC Directive) Regulations 2002 ("E-commerce Regulations") reg 17

[168] ibid reg 19

[169] Adi Robertson 'Reddit bans "deepfakes" AI porn communities' (*The Verge,* 7 Feb 2018) <https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban accessed> 11 Aug 2019.

[170] Samantha Cole, 'Twitter Is the Latest Platform to Ban AI-Generated Porn' (*Vice*, 6 Feb 2018) <https://www.vice.com/en_us/article/ywqgab/twitter-bans-deepfakes> accessed 11 August 2019; Samantha Cole, 'Pornhub Is Banning AI-Generated Fake Porn Videos, Says They're Nonconsensual' (*Vice*, 6 Feb 2018) <https://www.vice.com/en_us/article/zmwvdw/pornhub-bans-deepfakes> accessed 11 Aug 2019

[171] Cole 'Pornhub' (n 172)

[172] Samantha Cole, 'AI-Moderators Fighting AI-Generated Porn Is the Harbinger of the Fake News Apocalypse' (*Vice* 16 Feb 2018) https://www.vice.com/en_us/article/d3wd3k/gfycat-fighting-ai-porn-deepfakes-fake-news accessed 28 July 2019; Samantha Cole, Gfycat's AI Solution for Fighting Deepfakes Isn't Working (*Vice* 19 June 2018) <https://motherboard.vice.com/en_us/article/ywe4qw/gfycat-spotting-deepfakes-fake-ai-porn> accessed 28 July 2019

[173] Cole 'AI-Moderators' (n 174)

[174] ibid

© 2019 Danielle S. Van Lier

source material and rejects GIFs it determines are fake. [175] Another technology can search for the body and background.[176] Human moderators also review some flagged content.[177] While the AI is useful for catching celebrity deepfakes, it relies on publicly-available images and therefore has less utility when the subject does not have a significant online presence.[178] Other platforms have also use content filtration technology to combat harmful content, such as pornographic or jihadist content; expanding these tools to deepfakes should be possible.[179]

F. The Need for Comprehensive Legal Reform

Comprehensive legal reform relating to deepfakes and other forms of IBSA is needed. Control over one's sexual identity should be central to any legislation.[180] Amending existing revenge pornography laws to add deepfakes, without more comprehensive reform, would be inadequate. Notably, a growing chorus of commentators argue that revenge pornography laws are inadequate to address traditional photographs and videos, let alone deepfakes.[181] Comprehensive criminal law reform that recognizes IBSA as a sexual offense and effectively penalizes it is needed.[182]

Any new law should be drafted carefully and precisely, with clear definitions.[183] It should include reasonable exemptions, such as when the disclosure is in the public interest, to ensure free speech rights are protected.[184]

Laws covering IBSA should not require proof of the perpetrator's motive nor should it have any specific intent requirement as it is irrelevant to the harm victims experience. The only relevant motivation should be the victims' lack of consent to either the creation or disclosure, as consent for the former does not grant consent to the latter.[185] The *mens rea* should be knowledge based — whether the "defendant knowingly engaged in, or knowingly coerced another person to

---

[175] ibid

[176] Louise Matsakis, 'Artificial Intelligence Is Now Fighting Fake Porn' (*Wired* 14 Feb 2018_ https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/ accessed 11 August 2019

[177] ibid

[178] Cole 'AI-Moderators' (n 174)

[179] Jonathan Taplin, 'How to force 8Chan, Reddit and Others to Clean Up' *The New York Times* (New York 7 Aug 2019) <https://www.nytimes.com/2019/08/07/opinion/8chan-reddit-youtube-el-paso.html> accessed 11 Aug 2019

[180] Beyens (n 152) p 39

[181] ibid

[182] Kelly Johnson, Clare McGlynn and Erika Rackley 'Shattering Lives and Myths' <https://claremcglynn.files.wordpress.com/2019/06/shattering-lives-and-myths-final.pdf>

[183] Citron and Franks (n 94) p 386; Citron (n 83) p 1948

[184] Citron and Franks (n 94) p 388

[185] Beyens (n 152) 39; Johnson (n182); Citron (n 83) p 1947

© 2019 Danielle S. Van Lier

engage in" the activity — although some argue for a recklessness standard.[186] The law should not apply if the individual did not realize they were breaching the victims confidence.[187]

Criminal penalties should be commensurate to the conduct and should be sufficient to serve as a deterrent. Penalties should increase under certain circumstances, such as when the perpetrator both created and distributed the content.[188] Other exacerbating circumstances might include threatening conduct, including sextortion or rape threats, doxing, or an intent to harass or cause distress.[189]

In addition to these reforms, the law should mandate better training on IBSA for law enforcement and criminal justice personnel and greater resources to investigate IABSA crimes.[190] Civil legal aid should be expanded and funded to include legal advice and support for IBSA victims.[191]

More effective civil remedies, including privacy law reform, should accompany the criminal law reform.[192] For example, California is currently considering legislation that would create a private cause of action against someone who distributes an image or video that exposes the victim's "intimate body parts" or shows them engaged in a "sexual act."[193] The victim could seek, among other remedies, economic and noneconomic damages, punitive damages, and their attorney's fees and costs and would have the right to proceed pseudonymously.[194] Importantly, the legislation includes certain free speech-related exemptions, including for content used in the context of law enforcement investigations.[195] Additionally, it makes clear that inclusion of a disclaimer would not be a defense.[196]

Existing legal remedies are inadequate to address the potential harms deepfakes cause individuals. However, strengthening the laws, without more, will not be enough to prevent the spread of and harm from deepfakes, including to society, as discussed in the next sections.

---

[186] Citron (n 83) p 1947; Ryan 1058

[187] Citron and Franks (n 94) p 387

[188] Citron (n 83) p1949

[189] ibid p 1948

[190] Johnson (n182)

[191] Johnson (n182)

[192] Citron (n 83) p 1948

[193] California Assembly Bill AB-602 <http://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602> accessed 10 Aug 2019

[194] ibid

[195] ibid

[196] ibid

© 2019 Danielle S. Van Lier

5. Deepfakes as a threat to democracy

A truthfully informed society is critical to preventing and fighting authoritarianism.[197] Deepfakes are antithetical to truthful information and will exacerbate the disinformation wars that plague modern politics. They have potential to influence elections and to incite violence.

The last several years have seen widespread disinformation campaigns waged by foreign actors seeking to disrupt other nations' political process.[198] They frequently used social media posts to seed and amplify false stories.[199] The much-reported Russian interference in the last US presidential election is a notable example. Deepfakes have potential to enhance these campaigns.

A Political and geopolitical risks

There is growing fear that a well-timed deepfake could be used to tip an election, as was attempted with leaked and doctored documents in France and the US.[200] Deepfakes enhance the risk of "diplomacy manipulation" — the use of "technology to create the belief that an event has occurred to influence geopolitics."[201] Studies show that false stories about politics spread more effectively than other types of news stories, including those about business, terrorism, or science.[202] This bodes well for a malicious actor seeking to cause disruption.

Faked videos have long been used for political ends. US President Donald Trumpand his associates have frequently shared falsified videos to bolster the president or discredit those who disagree with him. In May, he shared a video of Democratic House Speaker Nancy Pelosi that was doctored to make her appear publicly drunk.[203] Associates of his shared altered videos of a journalist to justify suspending his credential to access press briefings.[204] These videos have been proven false and widely condemned, yet social media platforms refused to remove

---

[197] Hall (n 65) p 67

[198] Chesney and Citron, 'Deepfakes and the New Disinformation War' (n 16)

[199] ibid

[200] ibid

[201] Charlie Warzel, 'He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse' (*Buzzfeed* 11 Feb 2018) <https://www.buzzfeednews.com/article/charliewarzel/the-terrifying-future-of-fake-news> visited 15 Aug 2019

[202] Hall (n65) p 55

[203] Igor Derysh '"Fox & Friends" admits Trump supporters doctored Pelosi video that depicts her as drunk' (*Salon* 24 May 2019) https://www.salon.com/2019/05/24/fox-friends-admits-trump-supporters-doctored-pelosi-video-that-depicts-her-as-drunk/ accessed July 24, 2019

[204]

© 2019 Danielle S. Van Lier

them.[205] The videos, created using traditional editing techniques, illustrates how deepfakes may be used to influence politics, particularly in key western countries.

Deepfakes already may have been used to impact politics in developing nations. In the African nation of Gabon, after months of its president not appearing in public, a video featuring his New Year's address was released to allegations from political opponents that it was a deepfake.[206] The video sparked an attempted coup.[207] Experts who analyzed the video agreed that it had traits typical of deepfakes.[208] In Malaysia, a video in which a man appears to confess to sex with a local politician also may have been a deepfake.[209] Most coverage has focused on the potential harms to western countries; deepfakes may have even greater potential to disrupt developing nations.[210]

B Threats to public safety

Deepfakes can be used to enflame public sentiment in a way that potentially endangers public safety and security. This can be useful in the hands of non-state actors, such as terrorist groups, or others who might seek to incite violence.[211]

Following the March 2017 terror attack in London, a photograph of a Muslim woman was used to incite anti-Muslim sentiment.[212] The photograph was shared by a Twitter account later determined to be a Russian troll account. [213] It was picked up and shared by far right activists in an attempt to fuel anti-Muslim sentiment.[214] The photograph was real, but had been taken out of

---

[205] Drew Harwell, 'Facebook acknowledges Pelosi video is faked but declines to delete it Washington Post 24 May 2019 https://www.washingtonpost.com/technology/2019/05/24/facebook-acknowledges-pelosi-video-is-faked-declines-delete-it accessed July 24, 2019

[206] Ali Breland, 'The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink' (Mother Jones 15 Mar 2019) <https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/> accessed 10 Aug 2019

[207] ibid

[208] ibid

[209] Harwell (n 37)

[210] Breland (n 206)

[211] Chesney and Citron, 'Deepfakes and the New Disinformation War' (n 16)

[212] Law Commission (n3) par 11.93

[213] Caroline Mortimer, 'Man who posted image of Muslim woman "ignoring Westminster terror victims" was a Russian troll' *Independent* (London 14 Nov 2017) <https://www.independent.co.uk/news/uk/politics/man-muslim-woman-london-terror-attack-phone-russian-troll-identity-a8052961.html> accessed 13 Aug 2019

[214] ibid

© 2019 Danielle S. Van Lier

context.[215] A few days later, the true context emerge, but the bots and activists had already spread it among their followers; the damage was done.[216]

Those with malicious intent similarly can use deepfakes to incite violence or enflame tensions. Foreign states and non-state actors can create inflammatory content to stoke existing societal divisions and tensions. For example, a government could use a deepfake of a violent action by a protestor as an excuse to crack down on ongoing political protests. Or insurgent groups and terrorist organizations could use deepfakes in myriad ways to maximize impact on their target audience and aid in recruiting.[217]

C. Threats to businesses

Deepfakes pose a threat to businesses. They are becoming a new tool in fraud and social engineering attempts.[218] They also can used to spread disinformation that can damage a company's bottom line. Companies need to adjust their incident response plans to address the potential fallout from deepfake-based attacks. [219]

Deepfakes already have been used to attack businesses.[220] The attacks were an evolution of a "business email compromise" attack, traditionally committed through spoofed emails.[221] The attackers identified company employees authorized to make financial transactions, then used deepfake audio impersonating a senior executive to authorize payments to a business entity they controlled.[222] Attackers hid imperfections in the deepfake by adding background noise.[223] These threats are not limited to businesses — individuals can similarly be coerced into sending or transferring money to assist a "relative" they are led to believe is in need. And with the

---

[215] ibid

[216] Gianluca Mezzofiore 'That tweet trolling a Muslim woman during the Westminster attack was actually by a Russian bot' (Mashable 14 Nov 2017)< https://mashable.com/2017/11/14/troll-fake-muslim-picture-westminster-attack-russian-bot/> accessed 13 Aug 2019

[217] Chesney and Citron, 'Deepfakes and the New Disinformation War' (n 16)

[218] Bernard Warner, 'Fighting Deepfakes Gets Real' (*Fortune,* 24 Jul 2019) <https://fortune.com/2019/07/24/fighting-deepfakes-gets-real/> accessed 11 Aug 2019

[219] Joan Goodchild Black Hat 2019: Deepfakes Require a Rethink of Incident Response ITPro Today 7 Aug 2019 https://www.itprotoday.com/big-data/black-hat-2019-deepfakes-require-rethink-incident-response accessed 11 Aug 2019 See also Warner (n 218)

[220] Scott Ikeda, 'The Cutting Edge of AI Cyber Attacks: Deepfake Audio Used to Impersonate Senior Executives' (*CPO Magazine,* 18 Jul 2019) <https://www.cpomagazine.com/cyber-security/the-cutting-edge-of-ai-cyber-attacks-deepfake-audio-used-to-impersonate-senior-executives/> accessed 10 Aug 2019

[221] ibid

[222] ibid

[223] ibid

© 2019 Danielle S. Van Lier

growing use of voice authentication technologies banking and other personal accounts are put at risk.

Much as well-timed deepfakes have potential to impact elections, they may do the same to a company's earnings and reputation. Well-timed deepfakes depicting improper or illegal conduct by a CEO or promising charitable donations that are never made can devastate a company's reputation.[224] Likewise, a deepfake depicting a financial analyst reporting earnings or growth issues ahead of a critical moment in the company's business could prove financially devastating. Much like sexually-explicit deepfakes, the potential reputational harms put companies at risk of blackmail and extortion. Deepfakes can do long-lasting harm to a company that the current law cannot remedy.

D. Impact on journalism

Deepfakes pose serious challenges for news organizations to ensure they are not unwittingly aiding the spread of falsified content.[225] Journalists may hesitate to trust video or audio accounts of events they did not witness firsthand. Social media and handheld video technologies provide unprecedented access to live, breaking news that mainstream news organizations can leverage to cover events in places where they do not have staff on the ground. The rise of deepfakes will cast shadows over these "firsthand accounts," delaying or even preventing timely reporting by traditional media outlets while amplifying voices that do not exercise the same degree of content verification.

Some newsrooms approach third-party video content "in terms of worst-case scenarios," questioning whether and how the video or its context could have been manipulated.[226] In attempting to verify videos, journalists should consider "where, why and how it was shared," trace it back to its source and ask questions.[227] One journalist created a deepfake video and circulated it among colleagues as an experiment.[228] Those who knew of the project quickly spotted issues including with audio-video synchronization and other idiosyncrasies with the depicted person.[229] Those who were not aware of the project felt a sense of unease and noted sound issues, but could not identify what was wrong with the video.[230]

---

[224] Aviv Ovadya and Hal Bienstock 'Is Your Company Ready to Protect Its Reputation from Deep Fakes?' (*Harvard Business Review*, 8 Nov 2018) <https://hbr.org/2018/11/is-your-company-ready-to-protect-its-reputation-from-deep-fakes> accessed 15 Aug 2019

[225] Baker (n 9)

[226] ibid

[227] ibid

[228] ibid

[229] ibid

[230] ibid

© 2019 Danielle S. Van Lier

Training journalists to understand and detect manipulated content will be important to fighting it.[231] The Wall Street Journal, for example, launched an internal task force led by its Ethics & Standards and Research & Development teams, with staff specially trained in deepfake detection. [232] The committee "host[s] training seminars with reporters, developing newsroom guides, and collaborating with academic institutions… to identify ways technology can be used to combat this problem."[233] They recommend newsrooms combine research, strategic academic partnerships and training on the tools and technologies.[234]

There are multiple ways journalists can attempt to verify footage in this deepfake age:

- Attempt to get verification from the source who provided it. If the source is unknown, there are steps they can take to try to verify its provenance.

- Find old versions of the footage. Because deepfakes and other altered videos are often based on existing footage, it may be possible to locate that footage and determine if or how it was manipulated.

- Examine the footage for glitches. This may involve going frame-by-frame, but deepfakes and other forms of digital manipulation often leave noticeable artifacts detectable with careful review. Even audio leaves behind hints.[235]

Newsrooms should consider technical solutions that can be employed to identify synthetic content, although detection technology will need to continue evolving along with the content creation tools.[236] Training journalists to recognize how content can be faked and employing an appropriate verification framework to examine the material, its source, and the context in which it was shared, will be critical in this age of deepfakes.[237]

Deepfakes have not caused the type of crisis that many have feared, particularly in politics. Significant misinformation activities and campaigns targeted at major political events have not used deepfakes yet.[238] Deepfakes have not been without harm, and they may yet cause the

---

[231] Fillion (n 77)

[232] Francesco Marconi and Till Daldrup, 'How The Wall Street Journal Is Preparing Its Journalists To Detect Deepfakes' (NiemanLab 15 Nov 2018) <https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/> 9 Aug 2019

[233] ibid

[234] ibid

[235] ibid

[236] Baker (n 9)

[237] Baker (n 9)

[238] Brandom (n 82)

© 2019 Danielle S. Van Lier

type of harm feared, but their primary use thus far has been for "misogynist harassment rather than geopolitical intrigue."[239]

6. What other solutions might help fill in current gaps in the law?

Existing remedies are inadequate to address the growing threats posed by deepfakes. As with other forms of harmful online content and conduct, academia, government, industry, and the public will all have a role to play.[240] Several suggestions have been made throughout this paper; this section proposes additional approaches to tackling deepfakes.

A. Technology-based solutions

i Deepfake detection technologies

Researchers are actively working on technology to detect and defend against deepfakes. Detection tools will need to be able to work in a "trustless environment" — one in which details of the video's provenance may be impossible to trace — and will need to work quickly to match the speed at which content spreads online.[241]

One early tool identified deepfakes based on the subject not blinking or blinking unusually, caused by the use of data (typically photographs) that depicted the subkect with their eyes open.[242] Unsurprisingly, soon after this research was released, subsequent algorithms generated deepfakes that blinked, limiting the tool's effectiveness.[243] More recently, a related team trained an algorithm to detect "face warping" artifacts left behind when the deepfake is created.[244] Because deepfakes are created using two-dimensional images that are rotated, resized and distorted to fit video of a person that was captured moving in three dimensions, certain elements of the person's head and face are often misaligned in the deepfake video.[245]

These are just samples of the work being done to develop detection technology. There are numerous others under development, including one that detects subtle affectations when

---

[239] Brandom (n 82)

[240] Warner (n 221)

[241] Harwell (n 37)

[242] Yuezun Li, Ming-Ching Chang and Siwei Lyu. 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking' (Unpublished Paper, ArXiv 2018) <https://arxiv.org/abs/1806.02877v2> accessed 7 Aug 2019; Agarwal (n 37); Siwei Lyu; 'Detecting 'deepfake' videos in the blink of an eye (*The Conversation*, 29 Aug 2018 <https://theconversation.com/detecting-deepfake-videos-in-the-blink-of-an-eye-101072> accessed 7 August 2019

[243] Agarwal (n 37)

[244] Li, Y and Lyu S, 'Exposing DeepFake Videos By Detecting Face Warping Artifacts' (Unpublished paper, ArXiv 2018) <https://arxiv.org/abs/1811.00656> accessed 18 Aug 2019

[245] ibid

© 2019 Danielle S. Van Lier

someone speaks[246] and one that uses reverse image search similar to Gfycat's AI technology.[247] Professor Siwei Lyu, who has been at the forefront of this research, describes "the competition between generating and detecting fake videos is analogous to a chess game."[248] As quickly as researchers develop detection technology, deepfake creators adapt their algorithms and datasets. As such, detection technology can only provide a partial solution to the problem. Some researchers are therefore working on defensive tools to protect against the creation of deepfakes or to otherwise verify .

ii Preventative and authentication technologies

One way to limit the harm of deepfakes is for technology to make them harder to create. Other proposed solutions verify the authenticity of content or providing depicted individuals with an authenticated alibi. There is also the potential for platforms to block the upload of pre-identified content.

A research team developed a way to add digital noise to a photograph that can fool face detection algorithms while remaining invisible to the human eye.[249] They foresee an application that would automatically add this noise to photographs as they are uploaded to social media or online sites.[250] This would potentially reduce the dataset available to deepfake creators.

Similarly, "digital provenance" technologies can help by authenticating content at the time it is created.[251] It would embed digital watermarks or fingerprints in the content and the metadata would be logged using blockchain or similar technology.[252] Camera-makers Nikon and Canon have developed and offered their own image authentication kits, both of which were cracked soon after release, rendering them ineffective.[253] Researchers also are exploring authentication technology to embed digital watermarks that would remain even as an image is post-

[246] Agarwal (n 37)

[247] Spivak (n 117) p 354

[248] Lyu (n 242)

[249] Li, Y and others, 'Hiding Faces in Plain Sight: Disrupting AI Face Synthesis with Adversarial Perturbations' (Unpublished paper, ArXiv 2019) <https://arxiv.org/abs/1906.09288> accessed 18 Aug 2019

[250] ibid

[251] Citron Chesney

[252] Citron (n 83) p 1958

[253] Spivak (n 117) p 353

© 2019 Danielle S. Van Lier

processed.[254] To be effective, solutions like this would require website operators be willing to block or identify content that lacks this provenance.[255]

"Authentic alibi services" or "life-logging" might become an option for high-profile individuals who have a need to protect their public reputations. This technology tracks and logs every aspect of one's life – a significant amount of data to turn over to a third party.[256] It also raises the potential for unprecedented mass surveillance by for-profit entities.[257]

Facebook developed a program that allows victims to preemptively submit photos that might become the subject of nonconsensual distribution in order to block them in the future.[258] The photos are "hashed" and the data is then used to block the future uploads of the same photographs. The same technology might not be as effective for deepfakes, where the content being created is new rather than existing photographs or video. However, once the deepfake is released to the wild, and therefore can be hashed, it potentially can prevent future uploads on platforms that adopt the technology. It requires a large degree of trust in the platform as the victim is voluntarily submitting their sexual content.

B. Reform of the online platform safe harbors

Reform of the online platform safe harbors is necessary to incentivize platforms to take action on deepfakes and other toxic content. This cry became louder in the wake of multiple mass shootings in the US in which toxic online content played a role.[259]

The safe harbor laws are a creatures of an earlier age, created as a way to allow the nascent internet to grow by giving passive intermediaries room to operate without concern they would be liable for users' speech they could not monitor or control.[260] The internet is now mature and platforms like Facebook have a global user-base larger than any single country.[261] These platforms have become active content intermediaries with their algorithms determining what content users see.[262]

---

[254] Lily Hay Newman 'To Fight Deepfakes, Researchers Built a Smarter Camera' (*Wired*, 28 May 2019) <https://www.wired.com/story/detect-deepfakes-camera-watermark/> accessed 17 Aug 2019

[255] Spivak (n 117) p 354-55

[256] ibid

[257] ibid

[258] ibid

[259] Taplin (n 181)

[260] ibid

[261] Ezra Klein, 'Mark Zuckerberg on Facebook's hardest year, and what comes next' (*Vox* 2 Apr 2018) <https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge> accessed 17 Aug 2019 (estimating Facebook's user base at 2 billion)

[262] Taplin (n 181)

© 2019 Danielle S. Van Lier

Policymakers should revisit the platform safe harbors and consider limitations to address deepfakes and similar harmful content. Platforms that encourage and facilitate the spread of deepfakes, particularly pornographic deepfakes, should lose their immunity and should be accountable to victims in civil actions. Other platform operators' immunity should be conditioned on taking reasonable steps to limit use of their platforms to spread harmful content. Some examples may be automated detection tools and better content moderation, as well as promptly responding to users' removal requests.

In its Online Harms White Paper ("White Paper"), the UK has proposed a new regulatory framework for platforms.[263] It creates a new independent regulator (discussed in more detail below) and proposes a new statutory duty of care that requires companies to "take reasonable steps to keep users safe, and prevent other persons coming to harm as a direct consequence of activity on their services."[264] Companies will be required to respond in a manner "proportionate to the severity and scale of the harm" and requirements will vary for content that is illegal versus that which is legal but harmful.[265] New codes of practice will set forth legal duties the companies will be expected to fulfill.[266]

One simple preliminary solution, whether voluntary or mandated by law, is for platforms to update their terms of service to ban deepfakes and then to actually enforce those agreements.[267] They presently lack transparency regarding their standards- and content-related decisions, which should be mandated as part of any legislative fix.[268]

i. Free expression as a limiting factor

The right to freedom of expression, including the "freedom to hold opinions and to receive and impart information and ideas" is a cornerstone of democracy[269] This right is not wholly unfettered and may be balanced against other, as may be necessary in cases involving the dissemination of deepfakes. Nonetheless, it poses an impediment to platform regulation if it is likely to cause excessive censorship. Lawmakers need to ensure this right is adequately protected in any change to the platform exemptions.

C. An independent regulator for online harms

---

[263] Digital, Culture, Media & Sport and Home Department, *Online Harms White Paper* (White Paper, CP 57, 2019) ("White Paper")

[264] ibid paras 3.1, 3.3

[265] ibid paras 3.4 -3.5

[266] ibid para 3.6

[267] Hall (n 65) p 72

[268] ibid

[269] Human Rights Act 1998 Art 10

© 2019 Danielle S. Van Lier

The White Paper proposes creation of an independent regulator to oversee enforcement of new duties that will be imposed under its proposals.[270] The regulator will establish codes of practice, including establishing a framework for transparency, trust and accountability, and will assess the companies' compliance with those codes and their own policies.[271] It will also oversee "the implementation of user redress mechanisms," and will take actions to enforce compliance.[272] Additional responsibilities will also promote education and awareness to empower users and the development and use of safety technologies. Lastly, it will undertake and commission research to better understand "online harms and their impacts on individuals and society.[273]

D. Ethics-based solutions

AI is a rapidly developing field with a commitment to making a lot of its developments open-source. This openness is part of what has fueled the rapid advancements. But it is also what has allowed the technology's misuse. Google's commitment to building a "technology that could have an impact as profound as electricity" is what caused it to lose control of TensorFlow upon which most deepfakes applications are built.[274] Google's technology had a profound impact, but likely not the one it intended.

Companies like Google lack the incentive to limit their open-source offerings because they are good for business.[275] Open source code helps create a funnel for new AI development talent and it can lead to the development of projects that inspire additional projects.[276] But open-source technology also creates transparency, which has benefits.[277]

The researchers developing these technologies need to engage in self-reflection and consider potential misuse of their technology both in its development and before releasing it to the public. Google, Microsoft, and other large developers are vocal about ethical AI development.[278] Their own scientists sign ethical pledges and they have research ethics groups, but they do not provide ethical guidance to their users.[279] That defies logic, given the power of these tools. Providing ethics guidance to developers using their open-source tools will not stop those with malicious intent, but it will some to make more ethical development decisions.

---

[270] White Paper (n 265) paras 3.2, 5.1

[271] ibid para 5.2

[272] ibid

[273] ibid

[274] Gershgorn (n 7)

[275] ibid

[276] ibid

[277] Gershgorn (n 7)

[278] ibid

[279] ibid

© 2019 Danielle S. Van Lier

Some ethicists question the logic behind allowing the technology sector to regulate itself and propose asking larger, existential questions around ethics and regulation in technology.[280] Deepfakes are one part of a larger societal discussion.[281] Whether voluntary or legally mandated, "an industry-wide commitment to basic legal standards, significant regulation and technological ethics" may form a partial solution to addressing the harms of bad technology design.[282]

E. Education and user empowerment

The White Paper proposes development of an online literacy strategy "to empower users to manage their online safety."[283] The review found that there is insufficient information on media literacy, particularly for adults, and that work was needed to address issues including disinformation and online attacks on women.[284] The White Paper sets forth an education strategy to address this.[285] Similar recommendations in the context of disinformation recommend prioritizing funding for news literacy programs.[286] Individuals, especially young people, need to learn how to critically evaluate news sources, especially those on social media or digital news sites.[287] The UK is already engaging in educational efforts about online safety in school curricula.[288] The public also needs to take steps to protect itself, such as getting their news from a diversity of sources and being skeptical of news from non-traditional sources.[289]

Users also need to understand the role they play in the spread of disinformation and take steps to prevent its spread. These include: checking the content's source; verifying the information's veracity; and understanding their own biases.[290] Education regarding disinformation, including how to spot potential deepfakes, may not prevent their creation but it can help limit their spread.

---

[280] Matt Beard 'To Fix The Problem Of Deepfakes We Must Treat The Cause, Not The Symptoms *The Guardian* (London 24 Jul 2019) <https://www.theguardian.com/commentisfree/2019/jul/23/to-fix-the-problem-of-deepfakes-we-must-treat-the-cause-not-the-symptoms> accessed 11 Aug 2019

[281] ibid

[282] ibid

[283] White Paper (n 265) para 9.18

[284] ibid para 9.16

[285] ibid paras 9.17 – 9.20

[286] Darrell M West, 'How to combat fake news and disinformation' (*Brookings* 18 Dec 2017) https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/ accessed 18 Aug 2019

[287] ibid

[288] White Paper paras 9.5, 9.13

[289] West (n 289)

[290] Aly Colón, 'You are the new gatekeeper of the news' (*The Conversation* 7 Feb 2017) https://theconversation.com/you-are-the-new-gatekeeper-of-the-news-71862 accessed 18 Aug 2019

© 2019 Danielle S. Van Lier

F. Limiting online exposure

Individuals can potentially limit their individual risk of being depicted in a deepfake by limiting the number of photos they post online. This can be done through social media privacy settings or simply not sharing or posting photographs. But not all photographs are within an individual's control, particularly those in the public eye who cannot control photographs like those taken by journalists, paparazzi, or the public or videos in which they appear. It also will not help guard against deepfake videos created by an ex-intimate or other individual with access to photographs of the individual. And in today's information society, it is not practical to withdraw from online life.

G. Require disclaimers – low-hanging fruit

One of the simplest solutions is to require that any content made using deepfake or similar technology be clearly and conspicuously labeled as such. This could be implemented in a variety of ways, such as automatic watermarking built into the creation software or upon upload to a hosting website. Watermarks might help identify deepfake videos when first posted but they can be removed.

Disclaimers alone are inadequate, especially with regard to pornographic deepfakes. Some deepfake pornography sites already label themselves as featuring deepfakes and contain disclaimers. For example, a website that describes itself as the "best deepfake porn site dedicated to fake celebrity porn" has a small print disclaimer.[291] The disclaimer, which notes that "all content [on the site is] fake" is in barely-readable dark grey against the black background of each page.[292] The site is not amenable to complaints — if visitors "have any issues with the content of this site, [they are asked to] leave and" not use it."[293]

Even if one website has a disclaimer or otherwise makes clear that its content is fake, that does not prevent the videos from spreading beyond that site without the disclaimer. Another drawback to requiring disclaimers or watermarks is that it potentially would defeat a defamation claim. Additionally, legally mandating disclaimers sends the message that society sees the content as acceptable. Not all deepfakes are harmful, and mandatory disclaimers or watermarks might be a way to preserve free expression in connection with those forms of content. But it should not be considered adequate with regard to sexual deepfakes.

6. Conclusion

For nearly two years, journalists have been portending deepfake-driven doom, such as a major geopolitical, national security, or other world-impacting crisis. As of this writing, that has not

[291] MrDeepfakes.com <https://mrdeepfakes.com> (Author's warning: this site contains very graphic videos on several of its pages) visited 31 May 2019

[292] ibid

[293] ibid

© 2019 Danielle S. Van Lier

happened and the technology primarily has been used to create fake pornography, content that is harmful in its own right. But that does not mean the concern is not legitimate. As the technology improves and with the geopolitical climate changing, the potential for deepfakes-based disruption remains.

Deepfakes and the technology behind them are not inherently evil or dangerous. They represent a potentially revolutionary evolution in audio and video technology that can and will have many beneficial uses. Filmmaking, education, disability assistance, and numerous other fields will benefit from the technology. But despite its positive applications, the technology has a dark side. It is this dark side that has been at the forefront of public discourse. The power of audio and video will make deepfakes a potentially powerful tool of disinformation. As computing power and deepfakes technology improve, deepfakes are poised to become powerful disinformation tools. They might even cause us to question the truth.

Presently, deepfakes technology is being used primarily as a tool of misogynistic exploitation. Deepfakes represent the newest form of image-based sexual abuse, turning celebrities, ex-intimates, classmates, and nearly any woman into a an unwitting subject of pornography. It is in this area that the law has its greatest shortcomings. The law generally fails women in addressing IBSA, but it is particularly weak in the context of deepfakes where there are no criminal penalties and civil remedies are grossly inadequate. Comprehensive reform that recognizes the severity of the harm caused by deepfakes is needed to both criminal and civil law.

Deepfakes remain a potentially potent tool to disrupt geopolitics and to endanger public safety. They also present risks to business and others for fraud or reputational damage. A well-timed deepfake has potential to tip an election, to enflame public sentiment to violent ends, or to disrupt a business' earnings. Journalists need to play a role in limiting the spread of deepfakes while not hindering the timely dissemination of breaking news.

The fight against deepfakes will require a multi-pronged approach. Academia, government, news media, industry, and the public will all have important roles to play. There is no single solution to stave off the threat deepfakes pose to freedom and democracy. One thing is certain — the law is not prepared to respond to the threat.

© 2019 Danielle S. Van Lier