Deep fakes: a challenge of the post-truth era

Alina GIOSANU
National Institute for Research and Development in Informatics - ICI Bucharest
Mareșal Alexandru Averescu Bvd, Nr. 8-10, Bucharest, Romania
alina.giosanu@ici.ro

Abstract: The informational society poses a lot of new challenges, the phenomenon of disinformation being one of the most prominent maladies in our times. The rapid digitalization is changing the face of our social interactions and is also altering our manner to relate to reality. This article deals with the phenomenon of deep fake videos in connection with the digital era, its relation with the propagation of fake news, discussing the harmful consequences of its proliferation, without disregarding that this new technology has also beneficial applications.
Keywords: deep fake, deep learning, generative adversarial network, artificial intelligence, cyber threats, neural network, fake news, informational cascade, filter bubbles

INTRODUCTION

In the 21st century, sometimes referred to as the "post-truth era", we are facing new challenges related to the information society: the communications revolution has led to the accelerated spread of lies, misinformation and suspicious claims. Especially in the online environment, we are bombarded with more and more information that becomes increasingly difficult to manage, control and verify. We tend to give credit to digital contents generated by close groups, like our friends and acquaintances, investing them with a priori trust. This tendency has been studied by researchers (e.g. Seeman, M., 2015, 2017) under the name of digital/information tribalism, more precisely translated as virtual communities of people sharing a common interest, reciprocally associated through social media platforms or other online mechanisms. This is the main driver of the propagation of false information. In this article I deal with the phenomenon of deep fake videos that may be perceived as the future of fake news, and their cascade-like spread which often leads to harmful effects. The name "deep fake" derives from the concept of "deep learning" – a type of machine learning based on artificial neural networks, used to synthesize existing images and videos by means of combining and superimposing them onto source images or videos.

DEEP FAKE VIDEOS – PROCESS AND DEVELOPMENT

Deep fake videos are essentially either videos or audio recordings faking the authentic ones, often used maliciously, or at least having intended/unintended harmful consequences.

Deep learning-based techniques which are specific Artificial Intelligence (AI) algorithms, make possible for users to manipulate/edit the content. Similar technologies are behind very popular apps like Snapchat and its Face Swap feature, allowing to switch faces between users. Another app that uses this technology is the Russian Face App with its popular `aging` option and more recently, the controversial Chinese app called Zao. The latter enables users to upload photos of themselves and be integrated into popular TV-show and movie scenes by swapping places with celebrities like Leonardo DiCaprio or Marilyn Monroe (see Murphy, C. and Huang, Z., 2019). Nevertheless, the major concern regarding deep fake photos and videos regards their use to generate pornographic content (deep fake pornography). Another disputable popular software is FakeApp which allows its users to create face swapped videos without any technical knowledge. As technology is rapidly evolving, deep fakes are becoming more and more realistic to the point that it would become impossible to tell the difference between real and fake.

As psychologists long ago pointed out, humans are marked by an inherent tendency to "search for, interpret, favour, and recall information in a way that affirms one's prior beliefs or hypotheses" (e.g. Plous, Scott, 1993). This specific, inductive type of cognitive bias is known as the "confirmation bias". Simply put, we sometimes tend to see what we want to see instead of the real thing. This human bias is exploited by those who create deep fake videos with malicious intent. Besides personal entertainment, other motivations of those who alter videos in this manner range from monetised entertainment such as Youtube (e.g. DerpFake) and pornography (e.g. AdultDeepfake) to blackmail, revenge porn, harm to individual/organizations, sabotage, or even nation-state influence (information or hybrid warfare), distortion of democratic discourse, manipulation of elections, undermining public safety and the list could go on.

GAN-TECHNOLOGY

The process of creating a deep fake video is based on GAN (`generative adversarial network`) which is a machine learning technique invented in 2014 by Ian Goodfellow. Initially, GANs were used to algorithmically generate new data categories from existing sets of data. A GAN can look at millions of photos of human beings and then produce a new photo that approximates the photos without being an exact copy of none. A website entitled `This person does not exist` (https://thispersondoesnotexist.com/) shows the impressive deep learning technique at work. Each time you refresh the website page, a new photo depicting a human face as imagined by a GAN is generated. Also, GANs may `look` at different photos of a single existing person and then create a new photo that has is not yet taken of that subject. In the same manner, the generative network can be used to create new audio from existing ones, as well as texts. As we see, GANs are a complex, multi-purpose technology. Their aim is to synthesize artificial elements (e.g. images) that cannot be distinguished from authentic ones.

As shown below in figure 1, each GAN is composed of two neural networks. These networks are sets of algorithms, roughly modelled after the human brain, conceived to recognize patterns. One of the networks is a generator synthesizing new items, and the other one a discriminator collecting samples from the instructive set of data and from the other neural network's output, predicting if they are `real` or `fake`. Concrete, real-world data is interpreted through a sort of "machine perception" which classifies raw information. All the sensory data (texts, images, sounds etc.) must be translated into numerical patterns contained in vectors. Gradually, the generator becomes able to synthesize more and more realistic images by collecting data from the discriminator. The latter network is also progressively improving, through the multitude of comparisons of samples with authentic images, making it hard for the generator to deceive it.

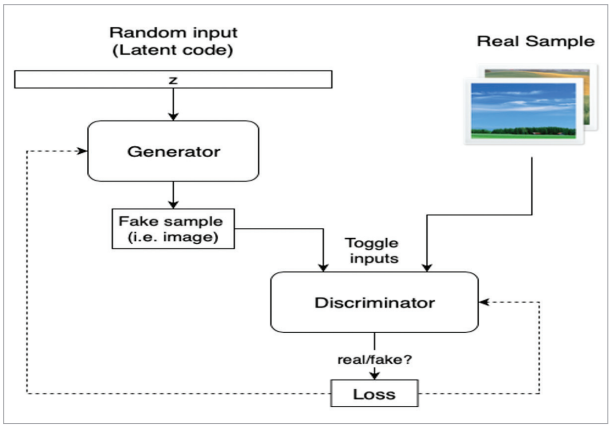

Fig. 1: *An overview of GANs (https://www.lyrn.ai/2018/12/26/a-style-based-generator-architecture-for-generative-adversarial-networks/)*

ProGAN (see Karras, Tero et al, 2018), an NVIDIA innovation from 2018, provides a solution for the lack of high-resolution in the generated large images, a problem that troubled the researchers. The novelty of ProGAN is "progressive training" which means that it begins by training the two neural networks with an extremely low-quality image, adding layers of higher resolution each time. As the resolution gradually increases, more and more details are learned over time. This specific type of training is also faster than the initial one but still not perfect – even if it manages to generate high-resolution images, its capacity to stably control different particular features is very reduced.

StyleGAN represents an improved version of the ProGan, focusing on the generator. This new alternative leads to "*an automatically learned, unsupervised separation of high-level attributes (e.g. pose and identity when trained on human faces {see Fig. 2}) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis*" (Karras, Tero et al, 2018).

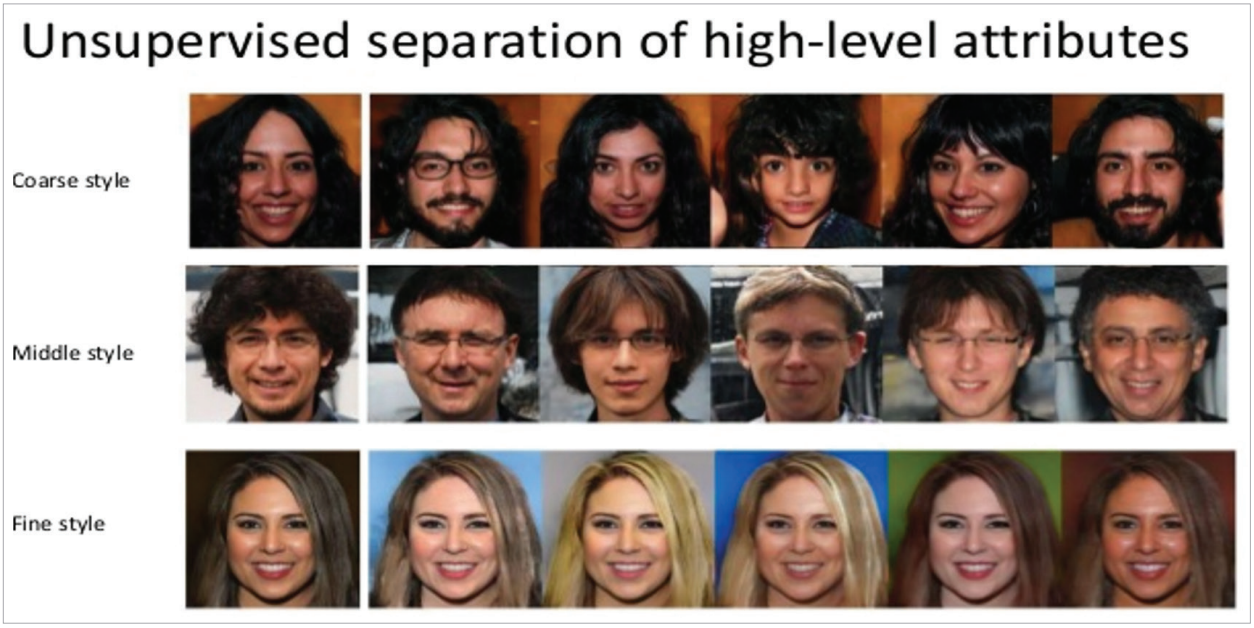

Fig. 2: *Style GAN unsupervised separation of high-level attributes (https://www.slideshare.net/ZhedongZheng1/style-gan)*

As we can see, the evolution of the GAN-based technique allows for increasingly more realistic generated images and, unfortunately, this represents also a high vulnerability. In the last years, cyber threats and crimes have become a priority on the majority of national political agendas. Decision makers struggle to regulate cyber security in order to address areas like the prohibition of cybercrime, the protection of critical infrastructures, cyber-attacks response or Internet governance, trying to find the best applicable strategies and policies without affecting the basic human freedom of expression. The nature of the Internet itself which allows a fluidity of identities and anonymity, making it very difficult to track the

author of a specific cybercrime. This adds to another bigger problem that is the difficulty to regulate the virtual domain. Sometimes, deep fake videos only aim at entertainment but the outcome of the rapid spread of such videos cannot be predicted. There are cases, usually involving politicians or celebrities, when deep fake videos have such an impact on the mind of a part of the viewers that they manipulate their thinking and making them adopt a specific political attitude. For example, Marco Rubio, a Republican senator who ran for president in 2016, made the following statement in relation to deep fake videos:

„In the old days, if you wanted to threaten the United States, you needed ten aircraft carriers, and nuclear weapons, and long-range missiles. Today, you just need access to our internet system, to our banking system, to our electrical grid and infrastructure, and increasingly, all you need is the ability to produce a very realistic fake video that could undermine our elections, that could throw our country into tremendous crisis internally and weaken us deeply." (Porup, J.M., 2019).

On the contrary, if you ask Tim Hwang, director at MIT Media Lab, he has a different perspective on the same subject:

„As dangerous as nuclear bombs? I don't think so. I think that certainly the demonstrations that we've seen are disturbing. I think they're concerning and they raise a lot of questions, but I'm skeptical they change the game in a way that a lot of people are suggesting." (Porup, J.M, 2019).

No matter how divergent the opinions about the impact of deep fakes, their rapid circulation in the online medium is a fact. The technological perfection of these fake videos is not even an essential factor for the rapid and ample propagation of false information. Let us look at three highly distributed, popular examples of deep fake videos:

• The highly suggestible human mind can very well be emotionally impacted by a low-quality deep fake video, or only a so-called "doctored video" like the popular one showing U.S. House speaker Nancy Pelosi, in which she sounds sluggish and slurred, as if she was inebriated. The sound of the footage was only slowed down to create this effect – a lot of people believed it was real at the time it appeared. That is why some called this video "cheapfake" or "shallowfake" (Ingram, M., 2019).

• Major examples of politicians and celebrities appearing in deep fake videos include the video

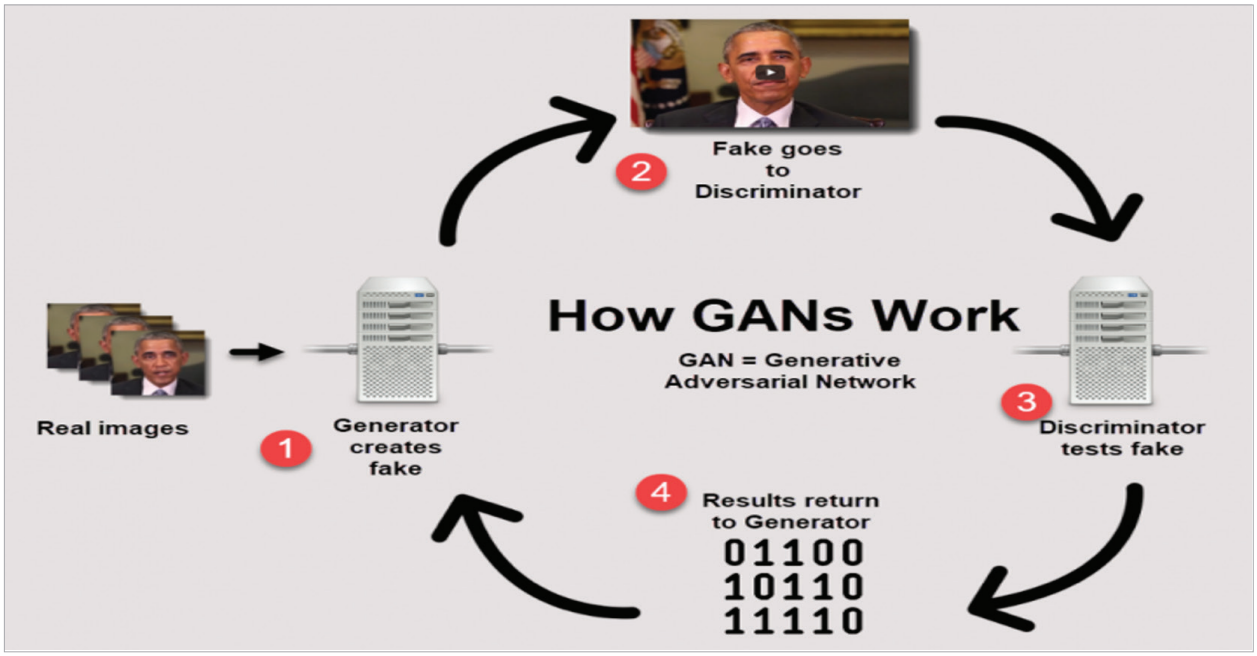

Fig. 3: *How GANs work (https://www.symantec.com/blogs/election-security/ai-generated-deep-fakes-why-its-next-front-election-security)*

in which former U.S. president Barack Obama appears to offensively refer to president Donald Trump. This famous deep fake was created by Oscar-winning TV-show and movie director, Jordan Peele.

• Another very popular case is the one where Facebook CEO, Mark Zuckerberg, appears to be stating that he has "total control of billions of people's stolen data, all their secrets, their lives, their futures".

Returning to the major importance of the human factor that contributes to a viral distribution of fake news and deep fake videos, their evolved "relatives", besides the human tendency to give credit to the opinions of close social groups members', a tendency that goes hand in hand with the algorithms of social networks that highlight the pieces of information you would most likely agree to/like/share, there is also a basic inclination to "*rely upon secondary information that doesn't come from any external source (...). This source is social information, or in other words: what we think other people are thinking*" (Chatfield, T., 2019). The sudden bursts of social information are named "infostorms" by researchers V. Hendricks and P. Hansen (2013):

"*Relying more and more on social information technologies or systems like these not only makes such sidetracking possible and more likely to occur, it also increases the numerical reach, if not the proportions, of the spreading of false beliefs and consequences thereof, intentional or nonintentional. When this happens we call the resulting phenomena infostorms*".

Researchers have also identified two phenomena that lead to massive online distribution (so-called "going viral"). Chesney & Citron (2018, p.10) note that "the interplay between cognitive heuristics (biases) and routine algorithmic practices make viral circulation possible". They identity two phenomena enabling this type of behavior: ***the dynamic of the information cascade***" and "***filter bubbles***" (Chesney & Citron, 2018, p.11).

The **information cascades** are resulting from the human tendency identified above, to rely on what other people know, even if it contradicts their own reason. This cascade-like phenomenon occurs when people give too much credit to what others know and stop paying enough attention to their own information. They will share the information, forwarding it, believing they have acknowledged something valuable and true. The **information cascade** is strengthened by the repetition of the same cycle on and on. The technology is a facilitator but the phenomenon is indebted more to our natural tendency to spread and promote negative and novel information (...) which "grabs our attention as human beings and causes us to want to share that information with others – we're attentive to novel threats and especially attentive to negative threats. (...). Coupled with our natural predisposition towards certain stimuli like sex, gossip, and violence, that tendency provides a welcome environment for harmful deep fakes." (Chesney & Citron, 2018).

The second identified phenomenon is the "**filter bubble**". The natural tendency to surround ourselves with information that supports our own opinions and beliefs is exacerbated by social media platforms by encouraging their users to re-share information. The algorithms of the social networks promote and highlight the most popular content, surrounding us with information from close groups. Because users share contents with which they agree, they are circled by information that supports their preexistent opinions – this phenomenon is called a "filter bubble".

As we can see, the viral propagation of falsehoods and decay of truth in general, as well as fake news/deep fakes in particular, are the result of the combination of common cognitive biases and social media capabilities. As Chesney & Citron (2018, p.14) efficiently summarize, "information cascades, natural attraction to negative and novel information, and filter bubbles provide an all-too-welcoming environment as deep-fake capacities mature and proliferate".

POSSIBLE SOLUTIONS

Let us explore three envisioned solutions that may prevent or at least attenuate the

harm caused by the malicious production and distribution of deep fakes: the improvement of public awareness, digital literacy and legal remedies.

1. The **IMPROVEMENT OF PUBLIC AWARENESS**: educational awareness regarding the ethical implications and the harm that can be easily inflicted to the person/s involved in the respective material. Education will lead to the prevention and reduction of massive distribution of false videos and news in general.

John Villasenor (2019) sees an important psychological effect of the deep fake phenomenon that we need to be aware of: "as we become more attuned to the existence of deep fakes, there is also a subsequent, corollary effect: they undermine our trust in all videos, including those that are genuine. Truth itself becomes elusive, because we can no longer be sure of what is real and what is not". This warning signal is to be linked with the post-truth era discourse.

2. Ethical education and awareness cannot be separated from a better, larger-scale **DIGITAL LITERACY**. Technological training may help a person discriminate an authentic video from a deep fake.

At a basic level, for example, several videos' characteristics might indicate that they are actually fakes:

- unusual lighting;
- discolorations of facial traits;
- blurred areas where the face meets the hair and the neck;
- blurred or disproportionate ears or teeth;
- changes in skin tone;
- double eyebrows/chins;
- face is getting blurry when it is partially concealed by a different object.

As we have seen, the human mind can be easily deceived partly because of the cognitive biases that are difficult to escape from. That is why, tech companies are working to develop AI-based tools capable of recognizing facial manipulation. In the end, this could lead to the creation of an end-product to assist consumers in detecting deep fake videos. For example, Adobe in partnership with researchers from the University of California Berkeley is working to create such a program. But we must keep in mind, as Tiffany Kelly (2019) points out, that "even the best fact-checking and identifying techniques are irrelevant if people think it's real and start to spread it on social media". There is always an interplay between technology and human perception: technology per se is neutral, we are the ones who render its value.

There are already available techniques that can help identify a deep fake but Matthew Stamm (in Chivers, T., 2019), assistant professor at Drexel University warns that "there's a lot of image and video authentication techniques that exist but one thing at which they all fail is at social media."

3. Besides an educational awareness and the development of apps capable of debunking deep fake videos which act as preventive solutions, **LEGAL REMEDIES** were discussed by law specialists as potential solutions that might limit the fraudulent use of these videos.

The problematics of a "flat ban" (Chesney & Citron, 2019, p.31), meaning a general prohibition of digital manipulation is rejected or at least considered "doubtful" by specialists. Chesney & Citron rightly note that "deep fakes exact significant harm in certain contexts but not all. A prohibition of deep fakes would prohibit routine modifications that improve clarity of digital content. It would chill experimentation in a diverse array of fields, from history and science to art and education".

Even if we agree that deep fakes should not be generally prohibited, there are particular situations in which their creators and distributors should be liable for the harms they inflict. A few obstacles (see Chesney & Citron, 2019) should also be taken into consideration here, of which maybe the most poignant is the attribution problem. In many cases, there is not enough data to be able to link a deep fake to his original creator (for example, if they use Tor, the so-called anonymity network, or other software facilitators for enabling anonymous communication). Another obstacle is related to the global, cross-border nature of the Internet. If there is a national legal process but the distribution exceeded the national borders, legal

action most likely would become inefficient. A third identified barrier is the difficulty to keep a process away from public interest. In most cases with deep fakes involved, the victim may not want to draw attention to the situation. As lawsuits attract publicity, filing the suit may aggravate the victim's harm.

In May 2019, legislators from New York have already proposed a revised Right of publicity bill, S5959, generally prohibiting the use of "a digital replica for purposes of trade in an expressive work" without the consent of the person. Under this law, it would become illegal to include a digital replica of, let's say, Keanu Reeves in a movie without his permission if it created "the reasonable impression" that he was genuinely performing. This law would specifically exempt newscasts and artistic works that do not trick the viewer into thinking they are watching the real person.

On June 13, the House Intelligence Committee convened to discuss the expanding threats to national security brought about by high-tech deep fake videos. When used as political weapons, consequences of deep fakes could potentially be disastrous. Federal law-makers introduced two bills regarding deep fakes, but no votes have yet been taken.

As we can see, legislation regarding deep fake falls into a grey area, as it poses a lot of difficulties. Further options are to be proposed and considered.

BENEFICIAL USES OF DEEP-FAKE TECHNOLOGY

We explored different facets of deep fake videos, especially focusing on their harmful effects. However, it is important to highlight once more that technology itself is neutral – it is in our power to orient it towards good or evil. In the end we would like to briefly note some of the beneficial uses of GAN-based technology, so we can have a full spectrum of its potential uses:

a. **Educational purposes**

Deep-fake technology provides a range of educational opportunities. To mention only one example, it offers the possibility to assemble videos of historical figures addressing directly to students in the classroom, creating an appealing alternative to readings and lectures.

b. **Applications in art**

The creative, artistic applications of deep-fakes are numerous. Video-artists can use them to create parodies, pastiches, critiques and satires of public persons and movie directors can choose to revive actors so they can be introduced in new films.

c. Self-expression

Deep-fake technology may be used to create an avatar-experience, like video games "that enable a person to have or perceive experiences that might otherwise be impossible, dangerous, or otherwise undesirable if pursued in person. The video game example underscores that the avatar scenario is not always a serious matter, and sometimes boils down to no more and no less than the pursuit of happiness." (Chesney and Citron, 2018) The Nintendo Wii ("Mii") avatars are perfectly illustrating the autonomy-related use of deep fakes.

CONCLUSIONS

As R.A. Wilson puts it, "we say seeing is believing, but actually, we are much better at believing than at seeing. In fact, we are seeing what we believe all the time and occasionally seeing what we can't believe". This quote clearly highlights the susceptibility of our biased mind in relation with the deep fake video phenomenon. In the so-called post-truth era (or even worse, in the "post-reality" era), we need to do our best to train our mind to resist to exaggerated credibility but in the same time pay attention not to become obsessively sceptical about each piece of information we encounter. This balance may be difficult to attain in a world marked by disinformation and fake news but nonetheless, we should aspire towards it. In this sense, deep fakes are truly a challenge of our contemporary, digital world.

ACKNOWLEDGMENT

This research was supported by Romanian Ministry of Research, project number 3N/2019.

REFERENCE LIST

Chatfield, Tom (2019), Why we believe fake news?, available at http://www.bbc.com/future/story/20190905-how-our-brains-get-overloaded-by-the-21st-century, accessed on 12th September 2019.

Chesney, Bobby and Citron, Danielle (2018), Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, 107 California Law Review (2019, Forthcoming); U of Texas Law, Public Law Research Paper No. 692; U of Maryland Legal Studies Research Paper.

Chivers, Tom (2019), What do we do about deepfake video?, availavle at https://www.theguardian.com/technology/2019/jun/23/what-do-we-do-about-deepfake-video-ai-facebook, accessed on 13th September 2019

Goodfellow, Ian et al (2016) , Deep learning: adaptive computation and machine learning, Cambridge, U.S.: MIT Press.

Hansen, Pelle G., Hendricks, Vincent F., Rendsvig, Rasmus K. (2013), Infostorms in Metaphilosophy Volume 44, Issue 3, LLC and Blackwell Publishing Ltd.

Ingram, Matthew (2019), Legislation aimed at stopping deepfakes is a bad idea available at https://www.cjr.org/analysis/legislation-deepfakes.php, accessed at 11th September 2019.

Karras, Terro et al (2017) , Progressive Growing of GANs for Improved Quality, Stability, and Variation available at https://arxiv.org/abs/1710.10196v3, accessed on 13th September.

Murphy, Colum and Huang, Zheping (2019), Social Media Users Entranced, Concerned by Chinese Face-Swapping Deepfake App, available at https://time.com/5668482/chinese-face-swap-app-zao-deep-fakes/, accessed on 14th September

Plous, Scott (1993), The psychology of judgment and decision making, Philadelphia: Temple University Press.

Porup, J. M. (2019), How and why deepfake videos work — and what is at risk available at https://www.csoonline.com/article/3293002/deepfake-videos-how-and-why-they-work.html, accessed at 10th September 2019.

Seeman, Michael (2017), Digital Tribalism – The Real Story About Fake News, available at http://www.ctrl-verlust.net/digital-tribalism-the-real-story-about-fake-news/, accessed on 15th September.

Seeman, Michael (2015), Digital Tailspin: Ten Rules for the Internet After Snowden, Amsterdam: Network Notebooks 09, Institute of Network Cultures.

Tiffany, Kelly (2019), Why it's harder to spot a deepfake once it goes viral, available at https://www.dailydot.com/unclick/detecting-deepfakes-sxsw-panel/, accessed on 13th September, 2019.

Villasenor, John (2019), Artificial intelligence, deepfakes, and the uncertain future of truth, available at:https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/, accessed on 11th September 2019.