The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability

Konstantin A. Pantserev

Abstract Contemporary psychological warfare has a number of instruments, including deepfakes, in which the human image is synthesized, based on AI algorithms. At first deepfakes appeared for entertainment. Special software based on artificial intelligence offers the opportunity to create clones that look, speak and act just like their templates. However, today the potential for deepfakes to be used maliciously is growing, whereby one creates a clone of a well-known figure and manipulates his or her words. This chapter analyses a wide range of examples of deepfakes in the modern world, as well as the Internet-services that generate them. It will also consider the possibility of using artificial intelligence to prevent their spread, as they constitute a serious threat to psychological security.

Keywords Psychological warfare · Deepfakes · Artificial intelligence · Deep learning · Fakes · Neural networks · Psychological security · Disinformation

1 Introduction

A new epoch is emerging in the history of world conflicts: an age of psychological warfare. Of course, psychological warfare as a phenomenon is not new, but today the development of information technologies (ICT) has rapidly increased its role in securing world dominance and geopolitical leadership, especially when even small military conflicts can now develop into global nuclear confrontation. Therefore, the contemporary global information and communication space has become the key battlefield and it is possible to say that we are now living in the era of global psychological warfare.

K. A. Pantserev (✉)
Saint-Petersburg State University, Universitetskaya nab. 7–9, Saint Petersburg, Russia
e-mail: k.pantserev@spbu.ru

H. Jahankhani et al. (eds.), *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*, Advanced Sciences and Technologies for Security Applications, https://doi.org/10.1007/978-3-030-35746-7_3

The key feature distinguishing such warfare from more conventional forms is the absence of any rules. Thus, whereas conventional military conflict should be conducted according to international conventions elaborated by the United Nations (and relatedly a number of international bodies exist with missions to supervise military conflict), psychological warfare can be conducted by any means. Consequently, mass disinformation and the production of fake news has become the key instrument of modern psychological warfare, following the principle of "if you can't convince somebody, misinform him or her".

AI-based deepfake technology is just one recent technological innovation that initially appeared as a form of entertainment but very quickly became a dual-use technology. This means that when used maliciously, deepfakes can threaten the psychological security of any state and international stability more broadly. In the following chapter the author will attempt to clarify the malicious use of deepfake technology.

2 From Fake News to Deepfakes

Contemporary information technologies are modernizing every day and offer considerable possibilities for users both in the field of business and entertainment. Thus, in order to provide some fun, there have appeared special web applications that produce fake news.

For example, the application *Fake TV News Maker*, which is positioned as the funniest generator of fake television news, offers the possibility to create TV news by using the interface of well-known news programmes. Using this application, one can create, change and distribute any news he or she wants. There is also the potential to replace famous speakers of TV news programmes who work for the largest information agencies and broadcasting companies, such as the BBC, ABC and CBS. *Fake TV News Maker* provides three categories for creativity: news, rumours and sport. Such news are carefully designed to appear like professional journalistic materials (what, when and where something happened), supplemented with photos that create the maximum realistic effect and thereby prevent fake and real news from being distinguished. The application also offers users the opportunity to share such "breaking news" with friends via e-mail, Bluetooth, Dropbox, Google Drive, Facebook, Flipboard, Google+, Hangouts, Linkedln, Picasa, SMS, Telegram, Twitter, WhatsApp and other social media and chats.

Another fake news generator, the *Fake Newspaper Maker*, provides opportunities for those who dream of seeing his or her news in newspapers with a large circulation around the world.

One application, *Fake News & Charts for iPad*, offers the possibility for those who are tired of fake news being distributed by unfair media and politicians to create his or her own "real" news.

Of course, one can find on the web a number of other fake news generators, such as *#FakeNews*, *Journalist CreativeBot*, *Fake News Creator*, *Fake Breaking News Maker*, *Make Fake News* and *Fake News Editor*.

For those who choose to prank, there is an application called *Text Now* that helps in the creation of fake text messages for "funny communication". This application offers the opportunity to create fake chats between any famous or even fictitious persons. There are also a number of other applications of such kind, like *Fake Message*, *Fake Text Message* and *Fake Call and Fake SMS*. And finally there is the *Fake News Generator*, the application which directly offers in one of its advertisements the chance to make fake news stories in order to fool friends and even the masses.

The appearance of all of these applications demonstrates that there is a demand for such production on the web. However, we suggest looking into the opposite site of all these jokes because they undermine the credibility and value of information distributed on the web and force us to seek an answer to the question regarding how we can trust such information: "How do we know that the computer is behaving as we expect it to or that an e-mail from our colleague is actually from that colleague?" [25].

Thus it becomes evident that the uncontrolled and spontaneous distribution of fake news represents a real and very serious threat to the psychological security of any country, and is capable of provoking psychological terrorist attacks that may cause fatal consequences.

Recent novelties in the field of artificial intelligence have also brought to life a new phenomenon, called deepfakes, which again initially appeared as a form of fun but today represent a serious challenge to psychological security.

3 Deepfakes: The Entity of the Technology

Deepfake, a portmanteau of "deep learning" and "fake", is a method of synthesizing a human image based on AI algorithms. Deepfake technology is based on two innovations in the field of machine learning: neural networks and generative adversarial networks, whose mission is to make deepfakes extremely realistic [6]. Neural networks represent a key element of machine learning technology.

> These are brain-inspired networks of interconnected layers of algorithms, called neurons, that feed data into each other, and which can be trained to carry out specific tasks by modifying the importance attributed to input data as it passes between the layers. During training of these neural networks, the weights attached to different inputs will continue to be varied until the output from the neural network is very close to what is desired, at which point the network will have 'learned' how to carry out a particular task [15].

Neural networks use the principle of the human brain's functioning: the more the human brain is exposed to information, the more accurately it can repeat it. Accordingly, the greater the number of examples downloaded to the neural network, the more carefully and correctly it can produce new examples [17].

In terms of deepfakes, the more video and audio data they download into the neural network, the more accurate the new audio and video will be, to the extent that it becomes impossible to determine whether this or that speech of this or that person is real or fake.

But, as Dack [7] argues, "neural networks are only half of the equation. Without generative adversarial networks, deep fakes would not be as realistic as they are". Generative adversarial networks were invented by Ian Goodfellow. He was formerly a Google researcher at the *Google Brain*, a special research team established in the early 2010s. This research team's mission was to make a breakthrough in the field of deep learning artificial intelligence. He is currently employed by *Apple*, another well-known corporation that conducts a research in the field of artificial intelligence. Goodfellow et al. [10] developed the idea of combining two neural networks together in order to make them compete with each other and hence improve the final product.

The principle of the work of generative adversarial networks is as follows. The first neural network, which is called the "generator", produces a new fake video or audio by copping the data set that has been downloaded. Next the original data set and the deepfake created by the first neural network are downloaded to the second neural network, which is called the "discriminator". Its mission is to distinguish a fake video from a true one [9]. If the discriminator is able to determine the fake video or audio, the generator tries to learn how the discriminator understood which video was fake and subsequently makes appropriate corrections. With each new iteration it becomes more and more difficult to distinguish a deepfake.

This presents the greatest problem. At present, when a deepfake is discovered, the appropriate correction is made and it will prove more difficult to discover the deepfake next time. Each detection of the deepfake improves it. Of course, researchers are working hard in order to improve methods of detecting deepfakes. For example, they pay attention to the frequency of the flicker of the image, natural micro-changes of the color of the face, or the irregularity of head or body movements, and so on. But all of those methods that help in identifying a deepfake today will fail to discover it in the future. Eventually, 1 day, there will appear a super-realistic fake video or audio that will be impossible to distinguish from a real one. One can use this for fun, just as a joke to fool one's friends, or (and much more seriously), to maliciously fool the masses by fake speeches of famous and influential persons such as politicians.

> Thanks to advances in artificial intelligence (AI) and computer-generated imagery (CGI) technology, over the coming decade it will become trivial to produce fake media of public figures and ordinary people saying and doing whatever hoaxers can dream of—something that will have immense and worrying implications for society [21].

The danger is that anybody can make any politician say whatever he or she wants, and then publishes this fake speech on YouTube or Facebook, on a fake website of the well-known mass media or on a fake social media profile of this or that politician. It will subsequently be shared by millions of people on social media. The fake video or audio can very quickly spread on the web and cause unexpected consequences,

such as by ending the political career of this or that person or even affecting the complexity of international relations between countries, possibly resulting in war.

> Progress in AI will enable new varieties of attacks. These attacks may use AI systems to complete certain tasks more successfully than any human could, or take advantage of vulnerabilities that AI systems have but humans do not [2].

Furthermore, some experts claim that "deep fakes don't need to be undetectable or even convincing to be believed and do damage. It is possible that the greatest threat posed by deep fakes lies not in the fake content itself, but in the mere possibility of their existence" [24]. This is because people believe in what they want to believe and they do not necessarily care whether a video is fake or real. Thus the main purpose is to feed people with the information they want to hear from the mouth of this or that politician. And this represents the greatest danger posed by deepfakes.

4 The Malicious Use of Deepfake Technology: From Fun to Psychological Warfare

As has been shown above, deepfakes can threaten personal, public and even national psychological security. It represents a very significant challenge of the contemporary digital age, as highlighted by many experts who are busy in the field of psychological and cyber security. For example, the company McAfee announced in March 2019 that it is today impossible to detect a change to a face using the naked eye. Thus, Steve Grobman, chief technology officer of McAfee, and Celeste Fralick, the company's chief data scientist, predicted in their keynote speech at the RSA Conference on cyber security issues in San Francisco that the use of new technologies by hackers is just a question of time.

Grobman and Fralick claimed in their speech that there is a special area in the field of cyber security called adversarial machine learning. Experts in this field are studying cyber-attacks to the machine learning classifiers. Grobman and Fralick argued that the method of replacing images represents a serious threat and can be used for the distortion of the work of the image classifier. As an example, they demonstrated one approach of how to deceive people using artificial intelligence: producing a real photo and carefully changing a small part of it. Through such minimal changes, a photo of penguins can consequently be interpreted by artificial intelligence as a frying pan. It is evident that false operation on a more serious scale can cause catastrophic consequences. Grobman stressed that deepfake technology represents a weapon that can be used for different purposes. It is impossible to prevent the malicious use of this technology, yet this is necessary to establishing a line of defense.

In order to prove that this threat is real, Grobman and Fralick presented a video in which Fralick's words were coming out from Grobman's face, even though Grobman never said them. Fralick concluded that this example "just shows one way

that AI and machine learning can be used to create massive chaos. It makes me think of all sorts of other ways in the social engineering realm that AI could be used by attackers, things like social engineering and phishing, where adversaries can now create automated targeted content" [27]. Every day such technology is improving.

Consequently, deepfake videos have become one of major challenges to the national security of any country, as they make people say or do things that they have never said or done. It is also necessary to point out that this technology is only 2 years old, and yet it has already made significant progress, with both the expert community and politicians starting to speak of the threat represented by deepfakes.

We'd like to remind that initially deepfake technology appeared as a fun in the "pornography industry where it referred to the process of inserting celebrities' faces into pornographic scenes" [7].

Such fake video clips first appeared in December 2017, when a user with the nickname *Deepfakes* published a pornographic video involving the famous actress Gal Gadot on the social media platform Reddit. Yet this video was a fake. In fact: he simply put the face of the Hollywood star onto the body of a pornographic actress with the aid of artificial intelligence. He "used TensorFlow, image search engines, social media websites and public video footage to insert someone else's face (it was a Gal Gadots' face—K.P) onto preexisting videos frame by frame" [13]. Since this video clip we have seen the widespread distribution of similar pornographic video clips involving other celebrities on the web. Users find an appropriate pornographic actress who meets all of their criteria and with the aid of the neural network change her face to that of the famous person of their choosing. The episode with Gal Gadot was just the beginning. Emma Watson, a British actress and model, was another victim of such fake pornographic films. The rapid growth of such videos has been clear. Thus, one can find on the web similar fake pornographic clips involving other famous women, whether Chloë Moretz, Jessica Alba, Scarlett Johansson or Maisie Williams.

Furthermore, fake videos in which a woman's face is changed into that of a man have also appeared, such as one of the actress Amy Adams being altered to that of the actor Nicolas Cage (to see this video follow the link: https://youtu.be/RdH7JoZZC2M). This "work" is actively used as an example of the possibilities of this technology, because it proves that it is possible to change a woman's face to a man's face. Nicolas Cages' face has subsequently appeared in many other domains, including movies such as *The Dark Knight Rises* and *Man of Steel* in which he never played. He has even replaced Sean Connery in the famous film *Dr. No*, Stephen Dillane in *Game of Thrones* and Harrison Ford as Indiana Jones in *Raiders of the Lost Ark*. It is difficult to determine why Nicolas Cage is so popular in such fake videos. Maybe people perceive his face type as the most suited for replacement. Perhaps the reason lies in the popularity of the actor. Indeed, Nicolas Cage "has become a meme of sorts in recent years in some corners of the internet. The deliberate incongruity of putting his image in unlikely places offers obvious opportunities for subversive humor" [22].

The considerable popularity and frightening realism of such fake video clips has forced Reddit, Twitter and even Pornohub to stop the distribution of video clips

created using AI-based deepfake technology. For example, the account of Deepfakes has been banned from Reddit. But the technology has already been successfully tested and today nobody can stop the continued distribution of fake videos that threaten any famous person, who must understand that 1 day he or she may be manipulated to appear in compromising videos. "Creating these deepfakes isn't difficult or expensive in light of the proliferation of A.I. software and the easy access to photos on social media sites like Facebook" [23].

Thus, the key feature of the contemporary digital age is the ease by which information technologies can be used. Indeed, the user does not need to know how neural networks work; he or she simply requires basic knowledge about computers and the Internet, find an appropriate application on the web, download it and have some fun.

Given the fact that AI technologies have become more and more accessible to ordinary users, coupled with easy access to photos and videos of former partners, colleagues and other people on Facebook and other social media, we are observing growing demand in programme tools that offer the possibility of creating fake videos.

One of these is *FakeApp*, a desktop application that can change faces in videos. It has been elaborated by the user *Deepfakes* on *Reddit* and is now distributed for free on the Internet. With the aid of neural networks, this application analyses the original faces of peoples and then tries to convert them into those of celebrities selected by the user. *FakeApp* is based on TensorFlow, Google's open-source platform for developing AI algorithms and the open-source library *Keras* [1]. It has a rather simple interface and offers detailed instructions on how to install and use the application. This simplification of complicated technology gives any user the opportunity to create fake videos and subsequently distribute them on the web via social media. The only thing he or she needs is an appropriate number of photos of this or that person so the application can learn and produce a highly realistic fake video clip. Here lies a great danger. As Chesney and Citron [6] argue:

> Imagine a video depicting the Israeli prime minister in private conversation with a colleague, seemingly revealing a plan to carry out a series of political assassinations in Tehran. Or an audio clip of Iranian officials planning a covert operation to kill Sunni leaders in a particular province of Iraq. Or a video showing an American general in Afghanistan burning a Koran. In a world already primed for violence, such recordings would have a powerful potential for incitement. Now imagine that these recordings could be faked using tools available to almost anyone with a laptop and access to the Internet—and that the resulting fakes are so convincing that they are impossible to distinguish from the real thing. Advances in digital technologies could soon make this nightmare a reality. Thanks to the rise of "deepfakes"—highly realistic and difficult-to-detect digital manipulations of audio or video—it is becoming easier than ever to portray someone saying or doing something he or she never said or did.

> Thus, the expert community marks out both a positive and negative side of this technological novelty. According to Francis Tseng, co-publisher of *New Inquiry*, an online magazine of cultural and literary criticism, on the positive side "deepfake technologies can bring to life new forms of art and even be used to create new genres of media ... And there can appear a whole culture of bootleg films produced in this way" [22].

But the use of deepfakes also has a negative side.

> As the technology improves, it will likely be used in more dangerous and antisocial ways. For example, it has the potential to turbo-charge fake news. When paired with technology that can synthesize real people's voices, apps such as FakeApp could make it extremely difficult for ordinary people to distinguish what's real from what's fake. And such technology could well be used to harass and blackmail people, putting them—virtually—in compromising situations . . . (and – K.P.) . . . lift cyberbullying to a whole new level [22].

In April 2018, a filmmaker named Jordan Peele sought to prove that this threat is a reality by using *FakeApp* to make a video in which former United States President Barak Obama insults the incumbent President Donald Trump. The voice of Obama is imitated by Peele himself. Of course, the production of this video required considerable time as well as specialists qualified in using special effects. In addition to *FakeApp*, *Adobe After Effects* was used for the editing of the video and its dynamic images, the development of the composition, animation and different effects. In total, it took about 56 h of automated processing of the video stream under the control of the specialist in special effects.

> Such technology should completely unnerve everyone of every political stripe, religion, sports affiliation, philosophical school and Jane Austin Book Club. Which is to say everyone everywhere. Because the capability works both ways, it's just as easy for your rivals to ruthlessly attack you with it as you them. It allows you to be "outed" by your enemies, for example, even if you have nothing to be outed for. It can now be done because it can. Any of us can be viciously slandered publicly by virtual "people" we don't know from Adam and who don't know us. Trash-talked and abused on the internet, TV and radio in every depraved manner imaginable, and unimaginable [26].

The rapid growth of such videos can be easily observed. Thus, a YouTube blogger with the nickname *Ctrl Shift Face*, who is familiar with deepfake technology, has introduced Jim Kerry in the role of Jack Terrence in Stanley Kubrick's film *The Shining*. The blogger manipulated the movie so that comic actor Kerry, in the role of Terrence (originally performed by Jack Nicolson) would punch a hole in the door with an axe and stick his head through it. This video clip has been published as open access on YouTube and is accessible at https://www.youtube.com/watch?time_continue=35&v=Dx59bskG8dc.

In discussing *Ctrl Shift Face*, it is necessary to mention that this YouTube blogger has rich experience in the production of deepfakes. Indeed, he has already changed Schwarzenegger to Stallone in *Terminator 2* (video clip accessible at https://www.youtube.com/watch?v=AQvCmQFScMA). But according to the evaluation of experts, his scene from *The Shining* is much more realistic. Kerry's face almost avoids "sliding" from the head of Nicholson, even when he makes sudden movements or covers his face with his hands. And what is next? YouTube followers of *Ctrl Shift Face* often request that he make a reverse replacement with Nicolson performing as Ace Ventura in the comedy film *Ace Ventura: Pet Detective*. The appearance of such video clips proves that deepfakes are not limited to erotic films. Therefore, if you want to see your favorite actor in roles originally performed by another person, the only thing you need is access to the appropriate AI algorithms.

Deepfakes also can be used to restore poor-quality videos. Indeed, a team of YouTube filmmakers from the channel the "Corridor" decided to modernize some of the worst scenes in the history of the film industry using special effects. For instance, they changed the scene in *The Mummy Returns* in which the King of Scorpions played by Dwayne Douglas Johnson first appears, the special effects of which have been greatly criticized. They thus used deepfake technology and the face of Dwayne Douglas Johnson to improve the scene. The final product as well as the filmmakers' explanations are available at https://youtu.be/KH1V6CHO1Jk.

A final example is that of American YouTubers *Sam* and *Niko*, who sought to create an amusing video with large numbers of views by producing a real film involving a fake Keanu Reeves, titled *Keanu Reeves Stops a Robbery*. It can be accessed at https://www.youtube.com/watch?time_continue=11&v=3dBiNGufIJw and has more than 1.4 million views. The story of this film is simple. Two men meet Keanu Reeves in the supermarket. One of them wants the actor to sign an autograph, but when they begin to talk, a robber enters the supermarket and tries to take money from the cashier. Keanu Reeves then attempts to stop the robbery without anybody suffering. He suggests that the robber take all of the money in his pockets, and on hearing the police arrive, he asks the robber to take his car and escape. But it is too late. The robber returns, followed by a cop. He kills the policeman and Keanu kills the robber. Keanu then gives the autograph to the man who had requested it, and leaves the supermarket. End of film. The value of this film is that it was not merely a scene from an existing movie: some YouTube bloggers decided to make a new film and invited Reuben Christopher Langdon, an American motion-capture and voice actor, to perform as Keanu Reeves. With the aid of deepfake technology, they altered Langdon's face to that of Reeves. The result turned out to be super-realistic.

All of the video clips described above were created in a short period of time in 2019. The extreme popularity of such videos enables us to conclude that deepfake technologies, introduced at the end of 2017, are rapidly spreading on the web and can serve different purposes. They can be used for fun by talented YouTube bloggers; they also can be used in the film industry when it is necessary to edit an old film, to improve a weak scene or to finish the film following the death of an actor. The film industry already has experience of using computer reconstructions of deceased actors. One of the most well-known examples comes from the film *Rogue One: A Star Wars Story*, in which British actor Peter Cushing appears in the role of Grand Moff Tarkin more than 20 years after his death in 1994! His image was reconstructed with the aid of a common gateway interface overlaid on a real actor, Guy Henry, who ensured a motion capture. The continued modernization of this technology will render this complicated technological process much easier.

However, such technologies can also be used maliciously, for example by compromising a famous person by making it appear that they are participating in intimate episodes or by spreading panic among the masses using speeches of different well-known figures. For example, in April 2018 there appeared a fake video involving Mark Zuckerberg, the Facebook founder, describing how he is going to close down Facebook and make its entire services toll.

This final point offers the opportunity to conclude that "[i]n our present age of misinformation, society will soon have to deal with deepfakes that can threaten national security. Consider a deepfake of President Trump announcing impending nuclear missile attack on North Korea" [14, p. 102].

Let's take another example. The Flemish Socialist Party created a fake video in which President Trump, during one of his speeches at the White House, calls for his country's exit from the Paris Climate Agreement and calls upon Belgium to follow the American example and withdraw as well (to see this video follow the link: https://www.facebook.com/Vlaamse.socialisten/videos/10155618434657151).

Towards the end of the video, Trump says, "We all know climate change is fake, just like this video". Of course, when evaluating this video one can see that Trump does not look as real and natural as he should, but this represents an example of how anybody can make a politician or any other person say anything he or she wants using deepfake technology. The rationale of the Flemish Socialist Party was to simply "start a public debate" and "draw attention to the necessity to act on climate change". At the conclusion of the video, it "calls on signing a petition that urges investing in renewable energies, electronic cars and public transport. The petition also calls on closing the Doel nuclear plant in Flanders". In order to create this video the "Flemish Socialist Party even used services of professional video studio checked with them that this is a legitimate procedure" [30]. This last statement demonstrates how today there is no law or any other power to prevent the further distribution of such fake videos. And what comes next?

According to Simon Chandler [4], a freelance technology journalist:

> Deepfakes are only the first step in a chain of technological developments that will have one distinct end: the creation of AI clones that look, speak, and act just like their templates. Using neural networks and deep learning programs, these clones will first exist in video and in virtual worlds. Whether you're knowingly involved or not, they'll provide exacting reproductions of your facial expressions, accent, speech mannerisms, body language, gestures, and movement, going beyond the simple transplanting of faces to offer comprehensive, multidimensional imitations.

This danger even forced the American TV company CNN to speak about the threat posed by deepfakes. The episode is entitled: "Rise of the "Deep Fake": Doctored Digital Videos Posing Threat to 2020, Security" (to see this episode please follow the link: https://edition.cnn.com/videos/tv/2019/01/29/lead-jake-tapper-dnt-deep-fakes-politics.cnn). On this episode is provided a fragment of the report to Congress made by Dan Coats, Director of National Intelligence, in which he offers a very real warning about media manipulation posing a major threat to US security. According to Coats, the country remains a superpower, but it has real competitors, with new technologies offering them opportunities to narrow the gap very significantly. One example is the famous fake video of Barak Obama, produced by Jordan Peele and described above. This video is frequently used to illustrate the possibilities of AI-based deepfake technology. Further details have been provided through an interview with Jeff M. Smith, a researcher from the National Center for Media Forensic of the University of Colorado at Denver, where he explains the basic principles of the work in deepfake technology.

Eileen Donahoe, a former US ambassador to the United Nations Human Rights Council, has also described the dangers of deepfakes in an interview with CNBC. She argued that deepfakes should be seen as the next generation of disinformation.

> In a year or two, as the algorithms continue improving, it's unclear whether the average person will even be able to discern authentic videos from fakes. At that point, even video evidence becomes questionable, and perhaps even unbelievable. In a world that already can't agree on simple facts, the future looks pretty terrifying [1].

Based on the facts stated above, one can conclude that today the malicious use of ICT is growing in unprecedented ways. We are already suffering from information garbage because the vast majority of information being disseminated on the web is produced by ordinary people who often use pseudonyms, rendering their identification very challenging. Now we come to the question of how can we trust information from anonymous or semi-anonymous sources? How can we distinguish true information from fake? Deepfakes render identification much more difficult or even almost impossible.

> Deep fakes represent a turning point in information warfare. They will increase the reach of fake news and decrease our connection to a shared understanding of facts. If people cannot trust what they see and hear with their own eyes and ears online, then they will choose what they want to believe [7].

Therefore, it is crucial to think about this threat very seriously and to elaborate real working instruments on how we can counteract the further distribution of deepfakes.

5 Countering the Deepfakes: From Theory to Practice

In discussing the necessity of tackling the distribution of deepfakes, one can argue that it is the key task of the state to protect its citizenry from toxic content. The state possesses different instruments with various levels of severity:

> ... online, by shutting down political websites or portals; offline, by arresting journalists, bloggers, activists, and citizens; by proxy, through controlling Internet service providers, forcing companies to shut down specific websites or denying access to disagreeable content; and, in the most extreme cases, shutting down access to entire online and mobile networks [16].

But the question is whether all of the measures mentioned above would be able to achieve positive results. Probably not. Once a website is closed down by the authorities, a number of new ones with almost the same content will appear. Furthermore, the arrests of popular bloggers, activists and leaders of oppositional movements can trigger mass protests. Therefore, more accurate and workable instruments need to be developed. In our opinion, solving the problem lies in two dimensions: a technological one and a legislative one.

As Gregory [12] argues, "in the category of technical solutions, many platforms, researches and startups are exploring using AI to detect and eliminate deepfakes.

There are also new innovations in video forensics that aim to improve our ability to track the authenticity and provenance of images and videos, such as *ProofMode* and *TruePic*, which aim to help journalists and individuals validate and self-authenticate media".

Today, computer science specialists are working hard to create appropriate algorithms capable of detecting deepfakes. Such algorithms could then be used by all major social media platforms, including Facebook, Twitter and YouTube, which are supposed to check all uploaded videos for deepfakes before they are made visible and accessible to other users.

But the problem is how to ensure deepfake detection with 100% probability. According to John Villasenor [29], "deepfake detection techniques will never be perfect. As a result, in the deepfakes arms race, even the best detection methods will often lag behind the most advanced creation methods". Consequently, elaboration of an effective and workable deepfake detection method constitutes a complicated task. Today there are a number of technical specialists working on this technological puzzle, proposing algorithms aimed at detecting deepfakes. Scholarly interest in this subject is also growing.

Indeed, Li Yuezun and Lyu Siwei from the Computer Science Department of the University at Albany, State University of New York have attempted to elaborate a method that would be able effectively distinguish AI-generated fake videos from real videos. In 2018 they proposed a method "based on detection of eye blinking in the videos, which is a physiological signal that is not well presented in the synthesized fake videos" [18]. However, this method cannot guarantee the detection of deepfakes completely because it is impossible to do so only based on the eye blinking. Therefore, Li and Lyu continued their research and in 2019 proposed an updated version of their method.

> Method is based on the observations that current DeepFake algorithm can only generate images of limited resolutions, which need to be further warped to match the original faces in the source video. Such transforms leave distinctive artifacts in the resulting DeepFake videos, and we show that they can be effectively captured by convolutional neural networks (CNNs). Compared to previous methods which use a large amount of real and DeepFake generated images to train CNN classifier, our method does not need DeepFake generated images as negative training examples since we target the artifacts in affine face warping as the distinctive feature to distinguish real and fake images [19].

David Guera and Edward Delp from Purdue University have proposed the use of neural networks to detect discrepancies between multiple frames in a video sequence that often occur as a result of face replacement. Their method "uses a convolutional neural network (CNN) to extract frame-level features. These features are then used to train a recurrent neural network (RNN) that learns to classify if a video has been subject to manipulation or not" [13].

A research team from the University of California, Department of Electrical and Computer Engineering, Naval Air Warfare Center Weapons Division, California, and Mayachitra Inc., Santa Barbara, has also developed a method for the detection and localization of fake images using resampling features and deep learning. In fact, they proposed two methods:

> In the first method, the Radon transform of resampling features are computed on overlapping image patches. Deep learning classifiers and a Gaussian conditional random field model are then used to create a heatmap. Tampered regions are located using a Random Walker segmentation method. In the second method, resampling features computed on overlapping image patches are passed through a Long short-term memory (LSTM) based network for classification and localization [3].

Both of those methods are aimed on the detection of digital manipulations such as scaling, rotation and splicing which are normally used in fake videos.

Social media platforms such as Twitter and Facebook are also involved in the elaboration of effective algorithms for the detection of deepfakes. For example, Antonia Woodford, the Facebook product manager, announced in September 2018 that Facebook has created a machine-learning model that would be able to detect "potentially bogus photos or video, and then sends these to its fact-checkers for review. Third-party fact-checking partners can use visual verification techniques, including reverse image searching and image metadata analysis to review the content" [28]. The deepfake content will be detected it is removed from the platform.

> Facebook intends to use its collection of reviewer ratings of photos and videos to improve the accuracy of its machine-learning model in detecting misinformation in these media formats. It's defined three types of misinformation in photos and video, including: manipulated or fabricated content; content that's presented out of context; and false claims in text or audio. Facebook offers a high-level overview of the difficulties identifying false information in image and video content compared to text, and some of the techniques it's using to overcome them. But overall the impression is that Facebook isn't close to having an automated system for detecting misinformation in video and photos at scale [28].

The latter statement by Tung demonstrates that Facebook remains far from elaborating a real workable algorithm for the detection of deepfakes. Furthermore, these methods and algorithms are mostly checking content that has been already published on social media. Therefore, before the deepfake is identified and banned, it may have been viewed by millions of people. This is why it is extremely important to think about how to identify deepfakes before they are made available to Internet users. This is the task of moderators of social media platforms, who must use appropriate AI-based algorithms in order to prevent the further distribution of deepfakes on the web. However, the problem is that today deepfake technology "is evolving so rapidly that as quickly as we can find ways to counter it, its creators can adapt it to make it more convincing" [8].

The final implementation of all of these algorithms aimed at detecting and blocking deepfakes can additionally face serious legislative problems because this field remains almost fully unregulated by lawmakers. This is why it is extremely important to think about the modernization of all national systems of laws in the field of ICT. The case of deepfakes proves that today information technologies are developing much more quickly than the national legislature, and large gaps exist in law in this sphere because almost every day new technological solutions that need to be regulated by law appear. The EU's measures, for example, "are still designed to target the disinformation of yesterday rather than that of tomorrow" [20].

This perspective is confirmed by Chesney and Citron [5], who state that today there is "no current criminal law or civil liability regime bans the creation or distribution of deep fakes. A threshold question is whether such a law would be normatively appealing and, if so, constitutionally permissible".

This is because this issue has a large range of legislative conundrums. "[T]here could be a lot of interesting [intellectual property] cases if amateur filmmakers start synthesizing films using the likenesses of celebrities and start profiting of that". [22].

Hence the first legislative problem detected is the need to stop the illegal distribution of images of persons on the web (we would like to remind the reader here that in order to create a deepfake, a large number of images of a person need to have been uploaded). This is a very difficult task because every day users upload millions of private images to different social media platforms not thinking that 1 day somebody will use them maliciously.

The next legislative puzzle lies in the nature of those synthesizing films. If one uses the face of this or that person in the creation of the fake film without his or her agreement, is it a crime? The national legislature of any country must find an answer to this question. Finally, a legislative solution on how to protect the private life of any person is required in the digital age.

But when elaborating anti-deepfake laws, it is necessary to consider the fact that sometimes deepfakes can have a positive impact. In other words, it is very important to try to find a solution regarding "how to distinguish malicious deepfakes from other usages for satire, entertainment and creativity, how to distinguish levels of computational manipulation" [11]. This is why it is first necessary to define the malicious use of deepfake technology. Only then it will become possible to obligate social media platforms to monitor, verify and remove toxic content.

However, any state adopting anti-deepfake laws can face a number of serious problems. From the first site it is evident that when countering fake videos, it is necessary to develop criminal laws regarding terrorist propaganda, inciting ethnic hatred, distributing knowingly fake information, using images of people without their permission, or invading their privacy. On the other hand, every person has civil instruments available when protecting his or her interests. Thus, he or she can sue for slander or for being portrayed in a false light. He or she can also sue for the use of his or her image without permission, seeking to prove that somebody else is benefiting from it. But when filing a lawsuit or elaborating anti-deepfake laws, it is necessary to consider how best to harmonize counteraction to deepfakes with freedom of speech and expression, which are considered fundamental human rights and are strictly protected by international conventions and the constitutions of the vast majority of countries. In the United States, for example, freedom of speech is protected by the First Amendment to the Constitution. Moreover, somebody can try to prove that deepfakes represent new forms of art and self-expression.

This would appear to constitute a serious obstacle to the adoption of anti-deepfake law. Yet at the same time, without such a law it seems unlikely that social media platforms will start to implement algorithms targeted at identifying and blocking fake video clips. Therefore, the problem of the mass distribution of deepfakes will remain highly significant for the foreseeable future.

6 Conclusion

To conclude it is necessary to highlight how today "we are on the verge of having neural networks that can create photo-realistic images or replicate someone's voice in a pitch-perfect fashion" [15]. This technology is known as deepfakes and represents a method of synthesising human images based on AI algorithms. The technology has existed for only 2 years but it has already achieved significant results. Now it produces super-realistic fake videos that are almost impossible to identify by the naked eye.

Initially the technology appeared as a form of fun in the pornographic industry, when somebody changed the face of an obscure porn actress to that of a celebrity. Thereafter a large number of similar videos have appeared on the web. *FakeApp*, a special desktop application that simplifies the process of creating fake videos, has also been developed, offering opportunities to anybody with basic computer skills to create deepfakes.

Today the use of AI-based deepfake technology is not limited to the pornographic industry and almost everybody can create deepfakes for entertainment, for business or for malicious use. Therefore, it is possible to recognize both the positive and malicious use of deepfakes.

Undoubtedly deepfake technology presents a wide range of possibilities. As Chesney and Citrion [5] argue, deepfakes can be used in education and offer the opportunity "to manufacture videos of historical figures speaking directly to students, giving an otherwise unappealing lecture a new lease on life. The technology opens the door to relatively cheap and accessible production of video content that alters existing films or shows, particularly on the audio track, to illustrate a pedagogical point". Deepfake technology can also be used in the film industry as it is now possible to use images of actors who have died to make new films or improve scenes of low quality. Finally, deepfakes open up a wide range of opportunities in the development of interactive television, as a user can change the actors involved. Thus, one must view deepfakes not only as a new technology but as a new form of art and self-expression.

However, deepfake technology can also be used maliciously and poses considerable danger both to personal and national security. It is necessary to remember that initially deepfakes were used in the pornographic industry to replace a porn actress' face with that of a celebrity, a "joke" that compromises the latter by her "participation" in the film.

In discussing the malicious use of deepfakes, it is necessary to recognize three levels of use: for individuals and organizations, for society and for the entire nation. According to Chesney and Citrion [5], deepfakes offer great opportunities for plotters to exploit and sabotage others in order to obtain financial or other benefits (this is true of the first level of malicious use of deepfakes, targeted at individuals and organizations). For example, "blackmailers might use deep fakes to extract something of value from people, even those who might normally have little or nothing to fear in this regard, who quite reasonably doubt their ability to debunk the

fakes persuasively, or who fear in any event that any debunking would fail to reach far and fast enough to prevent or undo the initial damage. In that case, victims might be forced to provide money, business secrets, or nude images or videos (a practice known as sextortion) to prevent the release of the deep fakes" [5].

The second level of malicious use is a threat to society. Fake videos offer major opportunities for plotters and terrorists to make politicians and other officials say and do things that they have never said or done.

> Fake videos could place them in meetings with spies or criminals, launching public outrage, criminal investigations, or both. Soldiers could be shown murdering innocent civilians in a war zone, precipitating waves of violence and even strategic harms to a war effort. A deep fake might falsely depict a white police officer shooting an unarmed black man while shouting racial epithets. A fake audio clip might "reveal" criminal behavior by a candidate on the eve of an election [5].

All of the examples mentioned above can cause deep political crises, end political careers, or even worse disturb relations between countries and thereby undermine international stability. This, the third level of malicious use, represents the greatest threat to national security.

Based on the above, it is possible to conclude that this technology bears a wide range of malicious use. But given that it can also be used for constructive purposes, one should think not how to completely stop the distribution of deepfakes, but rather to stop the distribution of toxic content.

In the opinion of this author, solving this conundrum will only be possible by combining technological and legislative methods. At the legislative level, it is necessary to elaborate a legal understanding of the malicious use of deepfakes and who (for example, service providers or social media platforms) should be responsible for detecting and blocking toxic content. At the same time, a workable AI-based algorithm aimed at quickly identifying and blocking deepfakes created for malicious purposes should be developed.

Given that this technology has only existed for 2 years, we do not expect a quick solution to this problem, although some algorithms aimed at identifying deepfakes have already been proposed and major social media platforms such as Facebook are conducting studies to try to block this content as soon as it is identified. Nevertheless, no country has yet adopted an anti-deepfake law. This is a considerable problem, because legal frameworks of deepfakes with clearly defined possibilities and cases of the legal use of such technology should be developed. Moreover, the main conundrum to be solved by lawmakers is ensuring a balance between forbidding the free distribution of deepfakes and protecting freedom of speech and self-expression, fundamental human rights protected by both international and national law. As has already been highlighted in this paper, deepfakes can be deemed a new form of art and self-expression. Thus, a simple banning of any fake video would grossly violate freedom of speech and self-expression.

Furthermore, the problem is that destructive elements that use deepfakes maliciously may also refer to the principal of the freedom of speech. This is why it is so crucial to solve this legal conundrum and distinguish the use of deepfakes for

malicious purposes from the legal use of such technology as a new form of art. Considering the fact that the technology to quickly identify deepfakes continues to lag behind in its development relative to that of their production, fake videos are likely to continue spreading widely on the web.

Until this occurs, people will continue to face the serious challenge of navigating around the information space and distinguishing true from fake information: even video scenes that look very realistic could in fact be fake. Thus, we come to the necessity of improving the basic communicative culture and information literacy of ordinary people.

This is a complicated process that requires systematic preventive work at all educational levels, from school to university. Scholars with different areas of expertise (for instance, psychology, social sciences and computer sciences) should be involved in this process. Together they should elaborate a curriculum for preventive work. Simultaneously, it seems extremely important that: (1) mass media informational and analytical publications emerge in which experts explain examples of the malicious use of information technologies and how one should navigate in the growing informational flow, distinguishing true information from false; (2) the organization of psychological master classes and seminars at schools, universities and other educational centers in order to improve people's communicative culture and information literacy; (3) conducting preventive conversations with parents, who should explain to their children from an early age the potential dangers posed by the Internet. Furthermore, in order to organize systematic work on this issue, it is extremely important that every state elaborate and adopt the "National Concept on the Organization of Preventive Work with Population", including the core principles of work aimed at improving people's communicative culture and information literacy. Only then will it become possible, if not to stop the further distribution of deepfakes for malicious purposes, but at least to neutralize their negative impacts.

Acknowledgements The author acknowledges Saint-Petersburg State University for research grant 26520757.

References

1. Browne R (2018) Anti-election meddling group makes A.I.-powered Trump impersonator to warn about 'deepfakes'. https://www.cnbc.com/2018/12/07/deepfake-ai-trump-impersonator-highlights-election-fake-news-threat.html. Accessed 19 July 2019
2. Brundage M et al (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf. Accessed 5 July 2019
3. Bunk J et al (2017) Detection and localization of image forgeries using resampling features and deep learning. https://arxiv.org/pdf/1707.00433.pdf. Accessed 30 July 2019
4. Chandler S (2018) Deepfakes 2.0: the terrifying future of AI and fake news. https://www.dailydot.com/debug/deepfakes-ai-clones-fake-news. Accessed 4 July 2019
5. Chesney R, Citron D (2018) Deep fakes: a looming challenge for privacy, democracy, and national security. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954. Accessed 26 July 2019

6. Chesney R, Citron D (2019) Deepfakes and the new disinformation war: the coming age of post truth geopolitics. Foreign Affairs, January/February. https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war. Accessed 26 July 2019
7. Dack S (2019) Deep fakes, fake news, and what comes next, 20 March. https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next. Accessed 12 July 2019
8. Fillion RM (2019) Fighting the reality of deepfakes. https://www.niemanlab.org/2018/12/fighting-the-reality-of-deepfakes. Accessed 30 July 2019
9. Fletcher J (2018) Deepfakes, artificial intelligence, and some kind of dystopia: the new faces of online post fact performance. Theatr J 70(4):455–471
10. Goodfellow IJ et al (2014) Generative adversarial networks, 10 June. https://arxiv.org/pdf/1406.2661.pdf. Accessed 12 July 2019
11. Gregory S (2018) Deepfakes and synthetic media: survey of solutions against malicious usages. https://blog.witness.org/2018/07/deepfakes-and-solutions. Accessed 23 July 2019
12. Gregory S (2019) "Deepfakes" are here, now what? https://internethealthreport.org/2019/deepfakes-are-here-now-what. Accessed 29 July 2019
13. Guera D, Delp E (2018) Deepfake video detection using recurrent neural networks. https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf. Accessed 4 July 2019
14. Harris D (2018) Deepfakes: false pornography is here and the law cannot protect you. Duke Law Technol Rev 17(1):99–128
15. Heath N (2018) What is AI? Everything you need to know about artificial intelligence: an executive guide to artificial intelligence, from machine learning and general AI to neural networks. https://www.zdnet.com/article/what-is-ai-everything-you-need-to-know-about-artificial-intelligence. Accessed 27 July 2019
16. Hussain M et al (2013) State power 2.0: authoritarian entrenchment and political engagement worldwide. Ashgate, Surrey
17. Karras T et al (2018) Progressive growing of GANs for improved quality, stability, and variation, 26 February. https://arxiv.org/pdf/1710.10196.pdf. Accessed 12 July 2019
18. Li Y, Lyu S (2018) In ictu oculi: exposing AI generated fake face videos by detecting eye blinking. https://arxiv.org/pdf/1806.02877.pdf. Accessed 26 July 2019
19. Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. https://arxiv.org/pdf/1811.00656.pdf. Accessed 23 July 2019
20. Meserole C, Polyakova A (2018) The West is ill-prepared for the wave of "deep fakes" that artificial intelligence could unleash. https://www.brookings.edu/blog/order-from-chaos/2018/05/25/the-west-is-ill-prepared-for-the-wave-of-deep-fakes-that-artificial-intelligence-could-unleash. Accessed 30 July 2019
21. Price R (2017) AI and CGI will transform information warfare, boost hoaxes, and escalate revenge porn. https://www.businessinsider.com/cgi-ai-fake-video-audio-news-hoaxes-information-warfare-revenge-porn-2017-8. Accessed 20 July 2019
22. Price R (2018) People are using creepy, cutting-edge AI technology to splice Nic Cage into every movie they can of. https://www.businessinsider.com/nicolas-cage-inserted-movies-fakeapp-ai-technology-2018-1?r=UK. Accessed 20 July 2019
23. Roberts J-J (2019) Fake porn videos are terrorizing women. Do we need a law to stop them? https://fortune.com/2019/01/15/deepfakes-law. Accessed 23 July 2019
24. Schwartz O (2018) You thought fake news was bad? Deep fakes are where truth goes to die. The Guardian, 12 November. https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth. Accessed 26 July 2019
25. Singer PW, Friedman A (2014) Cybersecurity and cyberwar: what everyone needs to know. Oxford University Press, Oxford
26. Snedeker R (2018) 'Deep fake': Obama didn't say what he just said. YIKES! 8 May. https://www.patheos.com/blogs/godzooks/2018/05/deep-fake-obama-video-yikes. Accessed 15 July 2019

27. Takahashi D (2019) McAfee shows how deepfakes can circumvent cybersecurity. https://venturebeat.com/2019/03/05/mcafee-shows-how-deep-fakes-can-circumvent-cybersecurity. Accessed 25 July 2019
28. Tung L (2018) Facebook's fact-checkers train AI to detect "deep fake" videos. https://www.zdnet.com/article/facebooks-fact-checkers-train-ai-to-detect-deep-fake-videos. Accessed 25 July 2019
29. Villasenor J (2019) Artificial intelligence, deepfakes, and the uncertain future of truth. https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth. Accessed 30 July 2019
30. Von der Burchard H (2018) Belgian socialist party circulates 'deep fake' Donald Trump video. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-videoE. Accessed 19 July 2019