

DIGITAL FORENSICS, DEEPFAKES, AND THE LEGAL PROCESS

By Agnes E. Venema and Zeno J. Geradts, PhD

Deepfakes are receiving much attention because of their potential for use and abuse, and the fear that they might influence election campaigns because of their ability to create a hard-to-detect alternative reality. A deepfake can be essentially described as the video equivalent of a photoshopped picture. However, the technology behind creating deepfakes is vast and complex. With their proliferation among the general public, deepfakes are also finding their way into the courtroom. The first cases of fraud by deepfake have already been reported on, and deepfakes of famous actresses being superimposed in porn videos are rampant. This article will discuss what we believe to be some of the current issues around deepfakes and the influence we expect them to have on judicial proceedings in the not-too-distant future.

Because deepfakes are becoming easier to create with standard software that can be downloaded from the Internet, their proliferation is increasing. Deepfake videos can be generated with software that can be used on phones, such as the app Zao (only available in China); open source toolboxes that exist on github, such as deepfacelab;¹ and for voice cloning, the toolbox Corintj, which can create a voice deepfake. Presently, home computers have GPUs (graphic processing units) that have enough computing power to make realistic deepfakes with a trained neural network that can be downloaded from the Internet. Previously, movie companies and a few other (state) actors

were the only ones to have enough computer power to create good deepfakes that were almost indistinguishable from real. However, as the software² for common use is getting better and easier to implement, and user interfaces are developed so that the software is accessible for more widespread use, even the average Windows user without much computer knowledge can start creating deepfakes.

In this article, we first look into methods for detection of deepfakes, which can be separated into manual and automatic detection. We then look at the impact of deepfakes on criminal and civil law, including on the role of juries. Finally, we conclude with future expectations and how they might impact the rule of law, society, and forensic science.

DEEPFAKE DETECTION: STATE OF THE ART

In this section, we review manual methods for the detection of deepfakes, as well as automatic detection of deepfakes.

Manual Detection

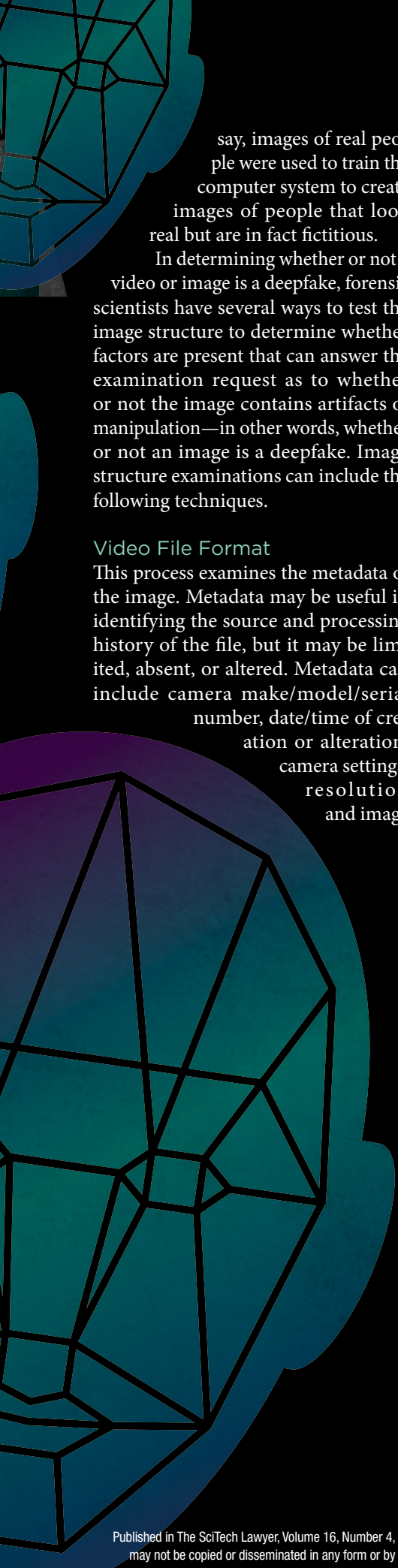
Deepfake detection techniques in forensic settings are based on image and video manipulation detection and are described in the best practices guide for image authentication of the Scientific Working Group on Digital Evidence (SWGDE).³ These are often rigorous and time-consuming methods because the forensic investigator needs to manually find artifacts of manipulation.

The most relevant aspects to manual detection of deepfakes are morphing and image creation. Morphing is a

combination of alteration and compositing. This is often used for deepfakes, either by using another face for the expressions and superimposing the expression, or even merging one face with another.

Image Creation⁴ through Artistic Means

An example of image creation through artistic means is the creation of virtual faces by training a deep neural network with a database of real faces. For example, the website thispersondoesnotexist.com is a compilation of still images created of persons that do not exist in real life but that are based on images of people who do exist. That is to



say, images of real people were used to train the computer system to create images of people that look real but are in fact fictitious.

In determining whether or not a video or image is a deepfake, forensic scientists have several ways to test the image structure to determine whether factors are present that can answer the examination request as to whether or not the image contains artifacts of manipulation—in other words, whether or not an image is a deepfake. Image structure examinations can include the following techniques.

Video File Format

This process examines the metadata of the image. Metadata may be useful in identifying the source and processing history of the file, but it may be limited, absent, or altered. Metadata can include camera make/model/serial number, date/time of creation or alteration, camera settings, resolution and image

size, Global Positioning System (GPS) coordinates/elevation, processing/image history, original file name, lens or flash information, frame rate, thumbnail information, etc. However, this process may be of limited use, as all the above data can be easily altered by saving the video in another format. An imagery file packaging analysis might be of use to discover if there are inconsistencies in how the file was stored.

Noise Analysis

A next step in the analysis involves an examination of noise within the image, and, if possible, the recording camera. Noise can be described as the tiny imperfections present in images that are the result of the sensors within the camera with which the images are created. If the camera is also in possession of the forensic scientists, they can use the camera to produce images to see if the noise—the imperfections embedded—on the new images are the same as those on the reference images.

The examination of noise can be done on different levels:

- Photo-Response Non-Uniformity (PRNU).⁵ This noise signature can be used to correlate images from the same source.
- Stochastic noise evaluation.⁶ This process can be used to show consistency between images from the same sensor manufacturer.

However, the results of these methods on their own should be used with caution, as these can also be spoofed⁷ (forged) or deleted with off-the-shelf software.

Assessment of Image Content

Image content examinations may include, but are not limited to, a review of the following: artifact features, artifacts in colors, breaks in compression, blocking or patterns, and unnatural motion in the video. In addition, physical aspects at the scene should be addressed, such as lighting, scale, and composition, as well as geographic

inconsistencies. When specifically looking at human subjects, one should look for artifacts in hair details, scars and bruises, creases, vein pattern, as well as skin contact and movement. Besides these, issues with focus, depth of field, sharpness and blurs, perspective, and noise, as well as lens distortion, should also be taken into account.

The examiner also needs to take into account the chain of evidence and if there was time for creating a deepfake. For example, if a video is from a CCTV-source captured in a closed circuit from a well-known manufacturer, it is unlikely that the video has been manipulated, unless there are signs of altered software at the level of the system itself. If videos have been downloaded from something like YouTube, the chain of evidence is less clear. All types of manipulation are possible between the time when the video was captured and the process of showing it on a computer.

Automated Detection

Many detection methods have already been developed for deepfake examination.⁸ Currently there are several databases, such as Face-Forensics++, that include thousands of examples of deepfakes versus non-deepfake videos, and most of these examples rely on Deep Neural Networks to find deepfakes. Databases such as Face-Forensics++ often report a detection accuracy of 91 to 95 percent.⁹ However, the real litmus test is how well the Deep Neural Networks that have been trained to detect deepfakes in these databases work in a real case. The databases include both original and manipulated videos, and, therefore, the ground truth can be known and determined with certainty if the Deep Neural Network correctly identified the deepfake or not. Furthermore, the system uses a limited set of data available to train its detection of deepfakes. A video from a real case falls outside the perimeters of the training data set, and it remains to be seen if the same level of accuracy can be achieved when such a video is introduced.

A major problem with the detection of deepfakes is that once a detection method is publicly known for any

deepfake generation method, it is possible that a creator can circumvent the detection method. For example, if vein patterns are used for detecting heart rate variations in deepfakes, the image can be altered in such a way that the heart rate variation in the whole image is equal.¹⁰ How easily one can create a nearly undetectable deepfake depends on the software that is available from the Internet.

DEEPAKES AND LAW

Deepfakes can affect both civil as well as criminal cases. However, the strong criminal and evidentiary focus best demonstrates the legal issues arising from the proliferation of deepfakes. Evidentiary standards in criminal cases are stricter than in civil cases, but that does not mean that deepfakes will not also impact, or give rise to, civil cases. However, whereas civil cases may more often evolve around the creation of a deepfake as the litigated act, in criminal cases the deepfakes can be used as a tool to commit another crime, such as fraud or child pornography. This digitalization of the tool is often referred to as cybercrime, but it may in fact just be regular criminal behavior committed by a digitalized means. The tool to digitalize ordinary criminal behavior, in this case, is the deepfake.

Prosecution

The prosecution of deepfakes is notoriously difficult.¹¹ Like much cyber- and cyber-enabled crime, the perpetrator is often unknown, as are his or her whereabouts. Even if the creator of the deepfake is known, prosecution can be difficult because most current legal frameworks are inadequate to deal with deepfakes.¹²

When the creation of a deepfake is considered to be the crime itself, copyright law and defamation laws appear to have the largest

degree of success.¹³ When invoking defamation law, however, the victim—especially when a public figure—may be required to demonstrate malice or negligence in order to bring a tort claim.¹⁴ The bar for bringing a defamation claim is not as high for persons without such a public profile, as they can bring a claim when falsehoods are circulated in a “merely” reckless or negligent manner.¹⁵ The more public the victim, therefore, the higher the threshold for being able to hold someone accountable for using his or her image to create a deepfake. To add insult to injury, the largest category of victims of deepfakes to date are famous actresses, female musicians, and other high-profile women who are featured in deepfake-generated porn.¹⁶ These individuals are considered to be public figures and thus the threshold, courtesy of the First Amendment, is therefore different than it would be for the average person.¹⁷ Furthermore, given the nature of many of these deepfakes, the adverse publicity that a public trial would generate will scare off many potential plaintiffs from filing a lawsuit, as for some the publicity may be worse than the crime.

Where an impersonation occurs with fraudulent intent, fraud- and forgery-related statutes may be invoked to successfully prosecute.¹⁸ As noted in the introduction, one of the most well-known cases that has been published to date relates to a UK-based energy company that is believed to have been defrauded because an employee thought he had spoken to his boss by phone, while the voice was in fact a voice clone deepfake.¹⁹ This audio clone was used to give the employees fraudulent orders over the phone, which then led to a wire transfer of approximately \$243,000 to bank accounts controlled by the criminals behind this scheme.²⁰

Not all deepfakes are created for illegal purposes. Deepfakes have often been used for entertainment or satirical value as well. In Italy, for example, a deepfake of the former prime minister was shown on a satirical TV show, swearing about his former colleagues and making vulgar gestures.²¹ When challenged, the creators evoked their

right to creative freedom and freedom of speech to defend their deepfake.²² The “victim,” the former prime minister, also took it that way; however, some prominent journalists and some members of the general public were thoroughly outraged. However, offensive discourse with a social or political purpose is considered protected discourse, especially in the United States.

What the Italian case demonstrates is that any new legislation regarding deepfake issues must be drafted in such a way as to not quell fundamental freedoms, such as speech and expression. Appropriate creative and potentially even therapeutic uses of deepfakes should be considered before unnecessarily strict legislation is passed.

Deepfake Defense

As opposed to the creation of the deepfake being the act under scrutiny, this section explores the role of audiovisual material as evidence. The risk of deepfakes can impact judicial procedures in two major ways: the deepfake defense, which claims that proof against a defendant is not authentic, and the introduction of a deepfake as evidence where it is not recognized to be inauthentic footage.

Most commentators think that the deepfake defense will debut in court in the foreseeable future if it has not done so already. The deepfake defense (also referred to as “the Liar’s Dividend”²³) is built around the premise that the audiovisual material introduced as evidence against the defendant is claimed to be fake or constructed. While shallow-fakes, videos that are original but have been manipulated in terms of slowing down or speeding up sections of footage, are relatively easy to debunk because original videos exist against which the shallow-fake can be compared, this is not possible for deepfakes.

The risks of the deepfake defense to the effective prosecution of a case are multiple. In some circumstances, audiovisual material may not be admitted as evidence because it may be considered to be “non-authentic” material. There is ample criminal case law in the US that allows imagery or (parts of) videos to be

admitted as evidence, so as long as their authenticity can be ascertained.²⁴ Expert testimony, such as that of digital forensic scientists, may be necessary to confirm the authenticity of the material that is to be submitted as evidence.²⁵ If developments in the field of creating deepfakes continue as fast as they have in the past, it is plausible that expert testimony and digital forensic tools may prove to be insufficient to verify the authenticity of evidence, leaving it to the prosecution to rely on other evidence or, in extreme cases, to drop the case all together.

If, however, audiovisual material is admitted as evidence, but a deepfake defense is raised, the burden of proof has now become one of proving a negative, rather than a positive, which is the opposite role many prosecutors will be used to. Instead of proving that the defendant committed the crime, the prosecution may now need to prove, beyond reasonable doubt, that the audiovisual material is, in fact, authentic and not digitally manipulated. However, proving a negative (in this instance) may turn out to be legally impossible. Current technology has great difficulty proving whether something is a deepfake, and expert testimony, as described above, may not be sufficient.²⁶

The opposite scenario, one where a deepfake is believed to be authentic and is entered as evidence, may also be a cause for concern and could lead to miscarriages of justice. A deepfake may be admitted as evidence of a crime without it being detected as nonauthentic. This could lead to someone being convicted on the basis of fabricated evidence, and in the worst-case scenario, it could lead to people being deliberately framed for crimes they never committed. As Maras and Alexandrou appropriately summarize, whereas before audiovisual evidence was often introduced to support witness testimony, we may see the reversal of the corroboration process.²⁷

Jury Instructions

When deepfakes make it into the courtroom, juries will have to be instructed on how to assess the evidence, including the deepfakes, in order to reach a verdict. In a world in which we have always

been inclined to believe what we see and where in fact we place more confidence in a visual representation of the facts than an oral description,²⁸ the jury instructions are likely to have to reflect this altered reality.

Audiovisual and photographic evidence are often digital and have been retrieved from a suspect's personal device. When deepfakes enter the world of evidence, jurors may be confronted with questions as to the chain of custody of the images and whether the evidence they see depicts the suspect in real life, or if the suspect's image was digitally manipulated into an existing video. The same scenario could be applied to the victim(s) that appear in (audio)visual evidence. It raises the question as to whether jurors could eventually see completely computer-generated evidence of a crime. If so, the jury instructions need to be very clear on what the criminal acts are, which of these acts can be proven beyond reasonable doubt, and what sentencing recommendations can be given for the different criminal acts. Furthermore, authenticity of the presented evidence will play a key role in arguing both the court case as well as the specific language of the jury instructions.

While the above hypothesis may seem extreme, variations of these issues may enter the courtroom in many forms in many types of cases. This is not a new issue; jurors have witnessed the evolution of biometric and technically complex evidence as they have unfolded over the past decades. A large body of academic literature exists regarding jury comprehension of complex evidence, such as DNA.

As certain crimes are committed now more often through digital means, the impact of the digital divide, the complex nature of the evidence, and the ability of jurors to understand it require new consideration. While no specific standards for detecting deepfakes exist, it may be worthwhile to inform juries of "generic" standards of digital and multimedia evidence, or the meaning of a digital chain of custody. And while oftentimes the deepfake will not be the only piece of evidence that will

be presented, deepfakes will inevitably become another aspect in the further digitalization of our lives, and by extension in the crimes that are committed and in the criminal justice system.

CONCLUSION

Future forensic scientists may be able to find evidence of the creation of a deepfake based on specific software that may be found on personal devices, such as laptop computers or mobile phones. The chain of evidence, however, is one of the key elements in this process. If a video is presented to the court without a good chain of evidence, questions as to the originality of the video may arise. If there are visible artifacts, it may be possible for forensic scientists to prove that a video is indeed a deepfake. However, if the deepfake has been processed in a professional manner, we expect neither algorithms nor humans to be able to distinguish the deepfake from the real photo, video, or audio fragment.

While prosecutors, judges, and jurors have been faced with rapid technological developments in the courtroom before, the potential impact of deepfakes on the judicial process is multifaceted. While the crime may be the creation of the deepfake itself, giving rise to claims of defamation or fraud, the law so far is ill-equipped to deal with a digital violation or laws written for an analog world. Furthermore, the introduction of a deepfake defense may lead to the prosecution needing to prove a negative, which could result in cases being dropped if the evidence is overwhelmingly audiovisual. Judges and juries alike need to be educated, and instructions given to juries need to be clear.

Given the trend to freely publish software, code, and training databases for the creation of deepfakes online, we expect to see the proliferation of deepfakes. The current trend of creating easy-to-use interfaces also makes the creation of deepfakes something that is accessible to the wider public, not just those with extensive computer or programming knowledge and topnotch hardware. We expect that the continuing publication of detection methods of

Continued on page 23

Deepfakes

Continued from page 17

deepfakes will eventually lead to deepfakes evolving quicker than the methods to detect them, as the “faking” software learns from the detection algorithms. In a world where we are accustomed to believe what our eyes see and our ears hear, the proliferation of deepfakes may lead us to have to reexamine our core belief systems, including those we rely upon in the courtroom.

Agnes E. Venema is a Marie Curie Research Fellow and PhD Candidate on the ESSENTIAL project at the Romanian “Mihai Viteazul” National Intelligence Academy and at the Department of Information Policy and Governance of the University of Malta. Her work focuses on the intersection of intelligence/national security, technology, and law. This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 722482. The opinions expressed in this article are those of the authors alone and not the funder.

Zeno J. Geradts, PhD, is a senior forensic scientist at Digital and Biometric Traces department of the Netherlands Forensic Institute of the Ministry of Justice and Security. He was president of the American Academy of Forensic Science from 2019 to 2020 and is one day a week a full professor at the University of Amsterdam in the field of forensic data science.

ENDNOTES

1. *Awesome Deepfakes*, GITHUB.COM, <https://github.com/aerophile/awesome-deepfakes> (last visited Apr. 15, 2020).
2. *Faceswap*, GITHUB.COM, <https://github.com/deepfakes/faceswap> (last visited Apr. 15, 2020).
3. SCI. WORKING GRP. ON DIGITAL EVIDENCE, SWGDE BEST PRACTICES FOR IMAGE AUTHENTICATION (July 11, 2018).
4. Shahroz Tariq et al., *Detecting Both Machine and Human Created Fake Face Images in the Wild*, in PROCEEDINGS OF THE 2ND INTERNATIONAL WORKSHOP ON MULTIMEDIA PRIVACY AND SECURITY—MPS ’18, at 81–87 (2d Int’l Workshop, Toronto, Canada, 2018), <https://doi.org/10.1145/3267357.3267367>.
5. B. van Werkhoven et al., *A Jungle Computing Approach to Common Image Source Identification in Large Collections of Images*, 27 DIGITAL INVESTIGATION 3 (Dec. 2018), <https://doi.org/10.1016/j.diin.2018.09.002>.
6. Ran Li et al., *Noise-Level Estimation Based Detection of Motion-Compensated Frame Interpolation in Video Sequences*, 77 MULTIMEDIA TOOLS & APPLICATIONS 663 (Jan. 2018), <https://doi.org/10.1007/s11042-016-4268-3>.
7. Sudipta Banerjee, Vahid Mirjalili & Arun Ross, *Spoofing PRNU Patterns of Iris Sensors While Preserving Iris Recognition* at 1 (2019 IEEE 5th Int’l Conf. on Identity, Security & Behavior Analysis (ISBA), Hyderabad, India, 2019), <http://arxiv.org/abs/1808.10765>.
8. Marwan Albahar & Jameel Almalki, *Deepfakes: Threats and Countermeasures Systematic Review*, 97 J. THEORETICAL & APPLIED INFO. TECH. 3242 (Nov. 30, 2019).
9. Pavel Korshunov & Sébastien Marcel, *Vulnerability Assessment and Detection of Deepfake Videos*, 12TH IAPR INT’L CONF. ON BIOMETRICS (ICB) (2019), <http://publications.idiap.ch/index.php/publications/show/4122>; Thanh Thi Nguyen et al., *Deep Learning for Deepfakes Creation and Detection*, ARXIV:1909.11573 [Cs, Eess] (Sept. 25, 2019), <http://arxiv.org/abs/1909.11573>.
10. Steven Fernandes et al., *Predicting Heart Rate Variations of Deepfake Videos Using Neural ODE*, in 2019 IEEE/CVF INT’L CONF. ON COMPUTER VISION WORKSHOP (ICCVW) 1721 (Seoul, S. Kor., 2019), <https://doi.org/10.1109/ICCVW.2019.00213>.
11. Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1788–803 (2019), https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=1640&context=faculty_scholarship.
12. *Id.* at 1792–803.
13. *Id.* at 1793–94.
14. *Id.* at 1793.
15. *Id.*
16. CENTRE FOR DATA ETHICS & INNOVATION (CDEI), DEEPFAKES AND AUDIO-VISUAL DISINFORMATION 10 (CDEI Snapshot Series, Sept. 2019).
17. Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99 (2019).
18. *See, e.g.*, Council of Europe, Convention on Cybercrime, tit. 2, Computer-Related Offenses, ETS No. 185 (July 1, 2004); Chesney & Citron, *supra* note 11, at 1802.
19. Jesse Damiani, *A Voice Deepfake Was Used to Scam a CEO out of \$243,000*, FORBES (Sept. 3, 2019), <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/#3937cfdd2241>.
20. *Id.*
21. *Deepfake Video of Former Italian PM Matteo Renzi Sparks Debate in Italy*, FRANCE 24: THE OBSERVERS (Aug. 10, 2019), <https://observers.france24.com/en/20191008-deepfake-video-former-italian-pm-matteo-renzi-sparks-debate-italy>.
22. *Id.*
23. Chesney & Citron, *supra* note 11, at 1785–86.
24. Marie-Helen Maras & Alex Alexandrou, *Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos*, 23 INT’L J. EVIDENCE & PROOF 4 (July 2019), <https://doi.org/10.1177/1365712718807226>.
25. *Id.* at 5.
26. *Id.*
27. *Id.*
28. *Id.* at 3.

