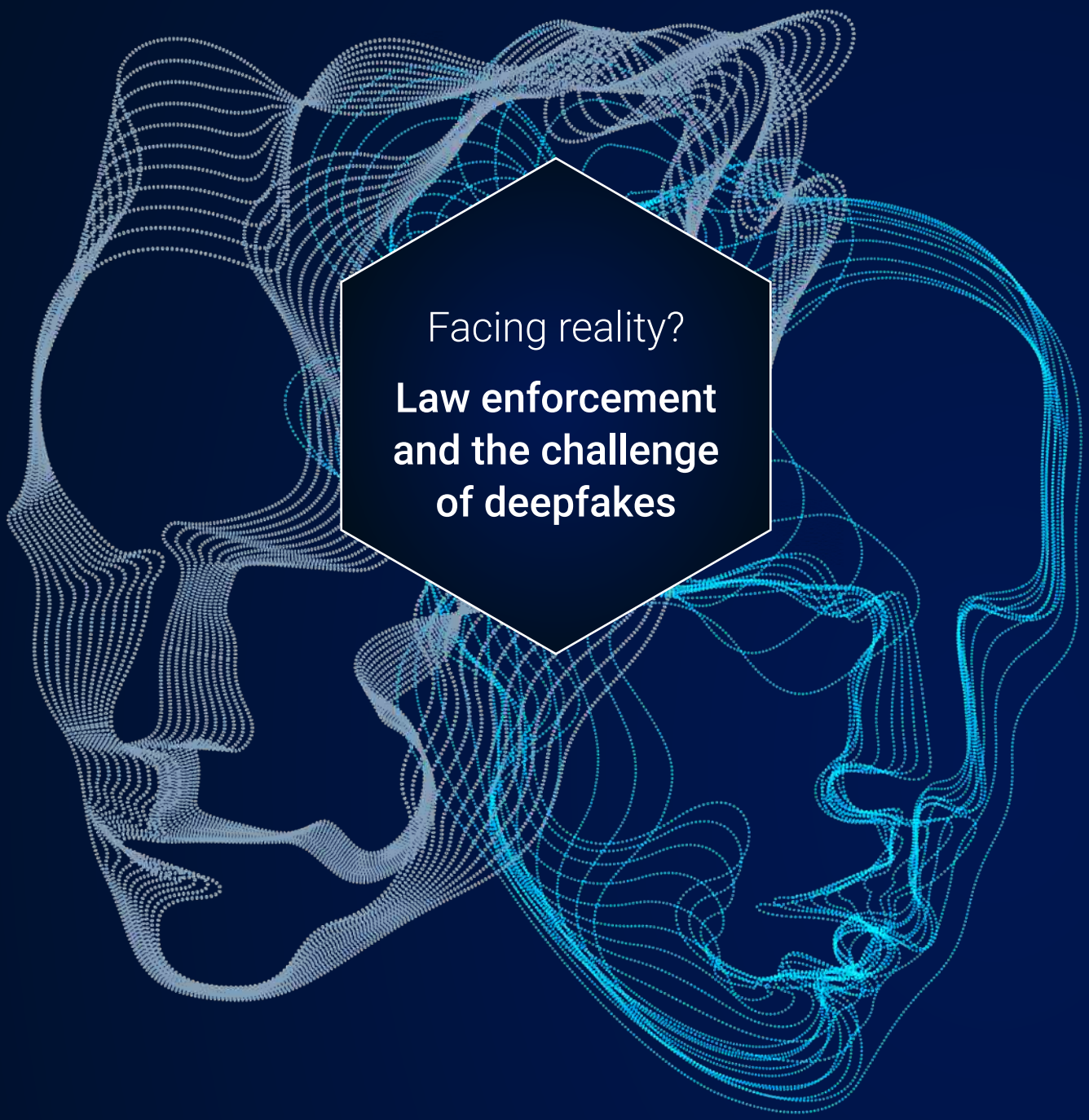
EUROPOL



Facing reality?

Law enforcement and the challenge of deepfakes

An Observatory Report from the Europol Innovation Lab

FACING REALITY? LAW ENFORCEMENT AND THE CHALLENGE OF DEEPFAKES
An Observatory Report from the Europol Innovation Lab

PDF | ISBN 978-92-95220-40-9 | ISSN 2600-5182 | DOI: 10.2813/08370 | QL-AS-22-001-EN-N

Luxembourg: Publications Office of the European Union, 2022

Cite this publication: Europol (2022), Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg.

This publication and more information on Europol are available on the Internet.

www.europol.europa.eu

Contents

Introduction

Today, threat actors are using disinformation campaigns and deepfake content to misinform the public about events, to influence politics and elections, to contribute to fraud, and to manipulate shareholders in a corporate context. Many organisations have now begun to see deepfakes as an even bigger potential risk than identity theft (for which deepfakes can also be used), especially now that most interactions have moved online since the COVID-19 pandemic. This concern is echoed by a recent report by University College London (UCL) that ranks deepfake technology as one of the biggest threats faced by society today.[1]

This poses a risk to EU citizens. Europol, as the criminal information hub for law enforcement organisations, will continue to play its part in supporting law enforcement authorities in the EU Member States to counter this threat.

This report presents the first published analysis of the Europol Innovation Lab's Observatory function, focusing on deepfakes, the technology behind them and their potential impact on law enforcement and EU citizens. Deepfake technology uses Artificial Intelligence to audio and audio-visual content. Deepfake technology can produce content that convincingly shows people saying or doing things they never did, or create personas that never existed in the first place.

To date, the Europol Innovation Lab has organised three strategic foresight activities with EU Member State law enforcement agencies and other experts. During strategic foresight activities conducted by the Europol Innovation Lab, over 80 law enforcement experts identified and analysed the trends and technologies they believed would impact their work until 2030. These sessions showed that one of the most worrying technological trends is the evolution and detection of deepfakes, as well as the need to address disinformation more generally. The findings in this report are the result of extensive desk research supported by research provided by partner organisations, expert consultation, and the strategic foresight activities.

Those workshops provided the initial input for this report. Furthermore, the findings are the result of extensive desk research supported by research provided by partner organisations, expert consultation and the strategic foresight activities conducted by the Europol Innovation Lab.

Strategic foresight and scenario methods offer a way to understand and prepare for the potential impact of new technologies on law enforcement. The Europol Innovation Lab's Observatory function monitors technological developments that are relevant for law enforcement and reports on the risks, threats and opportunities of these emerging technologies.

1 UCL – London's Global University, ''Deepfakes' ranked as most serious AI crime threat', https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat.

Understanding deepfakes

Disinformation is being spread with the intention to deceive. Tools of disinformation campaigns can include deepfakes, falsified photos, counterfeit websites and other information taken out of context to deceive the audience.[2]

In the original, strict sense, deepfakes are mostly disseminated with malicious intent, although they are now often used for positive applications too.[3] Experts estimate that as much as 90 %[4] of online content may be synthetically generated by 2026. Synthetic media refers to media generated or manipulated using artificial intelligence (AI). In most cases, synthetic media is generated for gaming, to improve services or to improve the quality of life, but the increase in synthetic media and improved technology has given rise to disinformation possibilities, including deepfakes.

Deepfakes were examined and discussed at great length in one of the Europol Innovation Lab's strategic foresight activities. Law enforcement experts who participated in these activities expressed concern about the consequences of disinformation, fake news and social media on political and social discourse. These trends are expected to become more pronounced as the supporting technologies, such as deepfakes, are becoming more sophisticated. Their impact on privacy and personal security will doubtless result in new categories of crime that will have to be policed. Participants were especially concerned about the weaponisation of social media and the impact of misinformation on public discourse and social cohesion.

On a daily basis, people trust their own perception to guide them and tell them what is real and what is not. This applies not only to people in their private lives, but also law enforcement officers trying to do their jobs. First-hand accounts are valued higher than second-hand versions of an event. Auditory and visual recordings of an event are often treated as a truthful account of an event. Photographs and videos are important intelligence for police work and evidence in court. But what if these media can be generated artificially, adapted to show events that never took place, to misrepresent events, or to distort the truth?

For instance, prior to the invasion of Ukraine by Russia in 2022, the United States revealed a Russian plot to use deepfake video to justify an invasion of Ukraine.[5] After the invasion happened, officials of the Ukrainian government warned that Russia might spread deepfakes that will show the Ukrainian president Volodymyr

2 Die Bundesregierung, 'What is disinformation?', accessed 15 March 2022, https://www.bundesregierung.de/breg-de/themen/umgang-mit-desinformation/disinformation-definition-1911048 .

3 ENLETS, 'SYNTHETIC REALITY & DEEP FAKES: IMPACT ON POLICE WORK', 2021, accessed on 15 March 2022, https://enlets.eu/wp-content/uploads/2021/11/Final-Synthetic-Reality-Deep-fakes-Impact-on-Police-Work-04.11.21.pdf.

4 Schick, Nina, Deepfakes: The Coming Infocalypse: What You Urgently Need To Know, Twelve, Hachette UK, 2020.

5 CBS News, 'U.S. reveals Russian plot to use fake video as pretense for Ukraine invasion', 2022, accessed on 10 March 2022, https://www.cbsnews.com/news/russia-disinformation-video-ukraine-invasion-united-states/.

Zelenskyy surrendering.[6] This fear appears to have become reality after hackers made a Ukrainian news website show a video of president Zelenskyy telling his soldiers to surrender.[7] At the time of writing much is still unclear about the video and it has not been verified to be a real deepfake or another fake, but it does show how the use of (deep)fakes are being used for disinformation purposes.

Examples like the one above show that this type of disinformation can be dangerous. Its aim is to intensify existing conflicts and debates, undermine trust in state-run institutions and stir up anger and emotions in general. The erosion of trust is likely to make the business of policing harder.

This challenge to policing is coupled with a public that seems relatively uninformed about the dangers of deepfakes. Despite their increasing prevalence at the time, research in 2019 showed almost 72% of people in a UK survey to be unaware of deepfakes and their impact.[8] This is particularly worrying as people might be unable to identify deepfakes (videos, photos, audios) since they are not aware of the existence of such virtual forgeries or how they work. The lack of understanding of the basics of this technology presents various challenges, some of which are relevant for law enforcement (such as disinformation and document fraud). Even more worrying results from recent experiments have shown that increasing awareness of deepfakes may not improve the chances for people to detect them.[9] Researchers are therefore expecting criminals to increase their use of deepfakes in the coming years.[10] This shows it is vital to understand the deepfake threat and prepare ourselves.

6 Metro, 'Ukraine warns Russia may deploy deepfakes of Volodmyr Zelensky surrendering', 2021, accessed on 15 March 2022, https://metro.co.uk/2022/03/04/ukraine-warns-russia-may-deploy-deepfakes-of-zelensky-surrendering-16217350.

7 National Public Radio, 'Deepfake video of Zelenskyy could be 'tip of the iceberg' in info war, experts warn', 2021, accessed on 17 March 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

8 iProov, 'Almost Three-Quarters of UK Public Unaware of Deepfake Threat, New Research', 2019, accessed 15 March 2022, https://www.iproov.com/press/uk-public-deepfake-threat.

9 Köbis, N.C. et al., 'Fooled twice: People cannot detect deepfakes but think they can', iScience, 24(11), 2021, accessed 15 March 2022, https://doi.org/10.1016/j.isci.2021.103364.

10 Recorded Future, Insikt Group, 'The Business of Fraud: Deepfakes, Fraud's Next Frontier', 2021.

The technology behind deepfakes

Deepfake technology uses the power of deep learning technology to audio and audio-visual content. Employed properly, these models can produce content that convincingly shows people saying or doing things they never did, or create people that never existed in the first place. The rise of the application of AI to generate deepfakes is already having, and will have, further implications for the way people treat recorded media. Here we discuss two core advancements behind deepfake technology, namely deep learning and generative adversarial networks, and how 5G technology may further enable the use of deepfakes.

Deep learning

Deep learning is a kind of machine learning where a computer analyses datasets to look for patterns with the help of neural networks. Machine learning is an application of AI where computers automatically improve through the use of data. Deep learning is a kind of machine learning that applies neural networks. These neural networks mimic the way our brains work to more effectively learn from the data provided. Deep learning technology, paired with the availability of large databases with material to train the generative models on, has allowed for rapid improvement of deepfake technology.

Deep learning algorithms use neural networks that mimic the brain's processes to find patterns in data.[11] Therefore, the availability of data is essential for a good deepfake system; it needs examples to learn what the result has to look like. It will try to discover patterns in the available data and thus extract what features are important and how these relate to each other. That will allow it to construct a complete and convincing picture. Depending on the quality of the available data and the factors the algorithm uses, the result may be more or less realistic.

Today, large datasets with labelled visual material are becoming freely available on the internet. These datasets are essential for the training of the machine learning algorithms needed to produce deepfakes. Creators of deepfakes can use these freely available datasets on the internet and avoid the time-consuming work of creating datasets themselves.

11 Code Academy, 'What Is Deep Learning?', 2021, accessed on 10 March 2022, https://www.codecademy.com/resources/blog/what-is-deep-learning/.

In one example from 2018, filmmaker Jordan Peele and BuzzFeed CEO Jordan Peretti created a deepfake video to warn the public about disinformation, specifically regarding the public's perception of political leaders. Peele and Peretti used free tools with the help of editing experts to overlay Peele's voice and mouth over a pre-existing video of Barack Obama. In the video, Obama allegedly said, "We are entering an era in which our enemies can make it look like anyone is saying anything, at any point in time. Even if they would never say those things."[12]

Source: Suwajanakorn, S. et al., 2017, 'Synthesizing Obama: learning lip sync from audio', ACM Transactions on Graphics, 36(4), accessed on 15 March 2022, https://dl.acm.org/doi/10.1145/3072959.3073640.

Generative Adversarial Networks (GAN)

A great leap in the quality and accessibility of deepfake technology was made by the adaptation of generative adversarial networks (GANs) as proposed in 2014 by Ian Goodfellow et al.[13] A GAN works with two competing models: a generative and a discriminating model. The generative model creates content based on the available training data, trying to capture the data as closely as possible, to create content that most closely mimics the examples in the training data. A discriminative model then tests the results of the generative model by assessing the probability the tested sample comes from the dataset rather than the generative model.

With the results from these tests, the models continuously improve until the generated content is just as likely to come from the generative model as the training data. This powerful method both simplifies the learning process, making it more accessible, and also improves the outcome by incorporating a mechanism designed to minimise the chance its product would be discriminated from authentic content.

When a new feature that may help discriminate between synthetic and authentic content is discovered, it allows for an easy

12 Ars Electronica, 'Obama Deep Fake', 2018, accessed on 10 March 2022, https://ars.electronica.art/center/en/obama-deep-fake/.

13 Goodfellow, I. et al, (2014), Generative Adversarial Nets (PDF). Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014). pp. 2672–2680.

incorporation of that feature. For example, people's eyes would not blink in early deepfake videos, making them relatively easy to detect.[14] Even though the training data for deepfake models included many pictures of people, these people generally did not blink in pictures. Adding more videos with people blinking to the database allowed both models to work together to produce people with blinking eyes, making the result more realistic and consequently harder to differentiate from authentic content.

Training data to create deepfakes may be applied in various ways for video and image deepfakes:

Face swap
Transfer the face of one person for that of the person in the video;

Attribute editing
Change characteristics of the person in the video, e.g. style or colour of the hair;

Face re-enactment
Transferring the facial expressions from the face of one person onto the person in the target video;

Fully synthetic material
Real material is used to train what people look like, but the resulting picture is entirely made up. See for example https://www.thispersondoesnotexist.com and https://generated.photos

Optimising these factors will improve the outcome. The more extensive the database and the more complex the algorithm becomes, the more computing power is necessary. Generating quality data requires a large volume and diversity of data with enough examples of similar but slightly different representations of the same characteristics to work. For example, if a database mostly contains pictures of white men with black hair, it will not perform too well on creating Asian women with blonde hair. As an increasing number and volume of databases are available, the quality and quantity of training data increases. This has allowed the models generating deepfakes to increase in sophistication.

Participants in the Innovation Lab's foresight activities noted how the roll-out of 5G would enhance connectivity and communication within law enforcement agencies (LEAs) and would strengthen the privacy and security of organisations and individuals alike. However, they noted that those same benefits would be leveraged by criminals to perpetrate their crimes. The additional bandwidth offered by new communication technologies, such as 5G, enables users to utilise the power of cloud computing to manipulate video streams in real time. Deepfake technologies can therefore be applied in videoconferencing settings, live-streaming video services and television.

14 GIZMODO, 'Most Deepfake Videos Have One Glaring Flaw', 2018, accessed on 10 March 2022, https://gizmodo.com/most-deepfake-videos-have-one-glaring-flaw-1826869949.

Deepfake technology's impact on crime

Participants of the foresight activities cited several trends that European LEAs should be sensitive to. Of note is crime as a service (CaaS), with criminals selling access to the tools, technologies and knowledge to facilitate cyber and cyber-enabled crime. CaaS is expected to evolve in parallel with current technologies, resulting in the automation of crimes such as hacking and adversarial machine learning and deepfakes. Indeed, participants flagged the tendency of criminal actors to become early adopters of new technologies. As a result, they are always one step ahead of law enforcement in their implementation, use and adaptation of these technologies.

The growing availability of disinformation and deepfakes will have a profound impact on the way people perceive authority and information media. With the increasing volume of deepfakes, trust in authorities and official facts is undermined. Experts fear this may lead to a situation where citizens no longer have a shared reality, or could create societal confusion about which information sources are reliable; a situation sometimes referred to as 'information apocalypse' or 'reality apathy'.[15]

This makes it essential to be aware of this manipulation and be prepared to deal with the phenomenon, so as to distinguish between benign and malicious use of this technology. The 'Malicious Uses and Abuses of Artificial Intelligence' report by Europol, TrendMicro and UNICRI[16] included a case study on this topic.

The report also shows that deepfake technology can facilitate various criminal activities, including:

- harassing or humiliating individuals online;
- perpetrating extortion and fraud;
- facilitating document fraud;
- falsifying online identities and fooling 'know your customer' mechanisms[17];
- non-consensual pornography;
- online child sexual exploitation;
- falsifying or manipulating electronic evidence for criminal justice investigations;
- disrupting financial markets;
- distributing disinformation and manipulating public opinion;
- supporting the narratives of extremist or terrorist groups;
- stoking social unrest and political polarisation.

15 The Guardian, 2018, accessed on 10 March 2022, 'An information apocalypse is coming. How can we protect ourselves?', https://www.theguardian.com/commentisfree/2018/mar/16/an-information-apocalypse-is-coming-how-can-we-protect-ourselves.

16 Europol, 'Malicious Uses and Abuses of Artificial Intelligence', 2020, accessed on 10 March 2022, https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence.

17 KYC stands for Know Your Customer and refers to the processes for identity verification and fraud risk assessment used by institutions.

Disinformation

Disinformation campaigns are operations to deliberately spread false information in order to deceive.[18] One major concern about this use is the ease of creating a fake emergency alert that warns of an impending attack. Another concern is the disruption of elections or other aspects of politics by releasing a fake audio or video recording of a candidate or other political figure. To illustrate this, the BBC created a video for the 2019 general election in the UK in which the candidates Boris Johnson and Jeremy Corbyn endorsed each other.[19] If this kind of manipulation successfully deceives a large enough part of the populace, this could have a serious impact on the outcome of an election.

Businesses are also at risk of being targets of disinformation, as deepfakes can be used to generate false information that could fool the public. For example, a threat actor could create a deepfake that makes it appear that a company's executive engaged in a controversial or illegal act. Certain deepfakes could be used for false advertising and disinformation, which could lead to bad publicity for a targeted company. Such applications of deepfakes could impact areas like stock market and company value as the public (stakeholders and shareholders, as well as consumers) may believe the deepfake and start selling their stocks or boycotting the company.

One example that shows the potential for criminal activities supported by deepfakes is the case where criminals used deepfake audio to impersonate the CEO of a company to make an employee transfer USD 35 million.[20] In this chapter/section of the report, we will look closer at four of the criminal uses of deepfakes that participants in the foresight activities identified.

Non-consensual pornography

In a December 2020 study, Sensity, an Amsterdam-based company that detects and tracks deepfakes online, found 85 047 deepfake videos on popular streaming websites, with the number doubling every 6 months.[21] In a previous September 2019 study, Sensity discovered that 96 % of the fake videos involved non-consensual pornography. To create this, one will overlay a victim's face onto the body of a pornography actor, making it appear that the victim is engaging in the act. In many situations, the victims of pornographic deepfakes are celebrities or high-profile individuals.

18 Marriam-Webster, 'Disinformation', accessed on 10 March 2022, https://www.merriam-webster.com/dictionary/disinformation

19 BBC News, 'The fake video where Johnson and Corbyn endorse each other', 2019, accessed on 10 March 2022, https://www.bbc.com/news/av/technology-50381728.

20 Forbes, 'Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find', 2021, accessed on 16 March 2022, https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions.

21 Sensity, 'How to Detect a Deepfake Online: Image Forensics and Analysis of Deepfake Videos', 2021, accessed on 10 March 2022, https://sensity.ai/blog/deepfake-detection/how-to-detect-a-deepfake/.

These videos are popular, having received approximately 134 million views at the time[22], and there are several pornographic sites that specifically produce pornographic celebrity deepfakes. Perpetrators often act anonymously, making crime attribution more difficult.

Document fraud

Passports are becoming increasingly hard to forge with modern fraud prevention measures. Synthetic media and digitally manipulated facial images present a new approach for document fraud. Using different methods and tools, it is possible to combine, or morph, the faces of the person the passport actually belongs to and the person(s) wanting to obtain a passport illegally. This method may increase the chance that the photo in a forged document passes any identity checks including those using automated means (facial recognition systems).[23]

The face in the middle of the image above is an example of a digitally manipulated facial image made using this 'morphing' method from the other two images. The images on the left and right are from The SiblingsDB, which contains different datasets depicting images of individuals related by sibling relationships. The subjects are voluntary students and employees of the Politecnico di Torino and their siblings, in the age range between 13 and 50.[24]

22 Government Technology, 'Deepfakes Are on the Rise — How Should Government Respond?', 2020, accessed on 10 March 2022, https://www.govtech.com/policy/deepfakes-are-on-the-rise-how-should-government-respond.html.

23 Robertson, D.J., Mungall, A., Watson, D.G. et al, 'Detecting morphed passport photos: a training and individual differences approach,' Cogn. Research 3, 27, 2018, accessed on 16 August 2021, https://doi.org/10.1186/s41235-018-0113-8.
MIT Technology Review, 'The hack that could make face recognition think someone else is you', 2020, accessed on 10 March 2022, https://www.technologyreview.com/2020/08/05/1006008/ai-face-recognition-hack-misidentifies-person.
Pikoulis, E.-V. et al., 'Face Morphing, a Modern Threat to Border Security: Recent Advances and Open Challenges', Applied Sciences, 2021, accessed 17 February 2022, at https://www.mdpi.com/2076-3417/11/7/3207.

24 T.F. Vieira, A. Bottino, A. Laurentini, M. De Simone, 'Detecting Siblings in Image Pairs', The Visual Computer, 2014, vol 30, issue 12, p. 1333-1345, doi: 10.1007/s00371-013-0884-3

This kind of approach to fraud can be applied to any other type of digital identity check that requires visual authentication. It greatly undermines identity verification procedures since there is no reliable way to detect this kind of attack.[25, 26, 27, 28]
Document fraud is a facilitator of other crimes like illegal immigration, trafficking in human beings, selling of various illegal goods, and terrorism, as perpetrators often use fake IDs to travel to their target locations. Deepfake technology might amplify the risk for advanced document fraud by organised crime groups.

In practice, the robustness of any identification process will depend on the process as a whole, and not only its visual step(s). However, a higher quality synthetic image will make a forged document more likely to pass the check of a visual identification step in the process. In general, the prospect of a successful document fraud attempt depends on quality and context of the deepfake used. The quality of the deepfake is largely dependent on available data and processing power, which is beyond the control of the identification process. The context in which the deepfake is applied is partially determined by the process however, providing opportunities to limit the success of just using a good deepfake.

Deepfake as a service

Just like many other new technologies, deepfakes are still used mainly by proficient engineers and research parties. However, deepfake capabilities are becoming more accessible for the masses through deepfake apps and websites. There are special marketplaces on which users or potential buyers can post requests for deepfake videos (for example, requests for non-consensual pornography). The increased demand for deepfakes has also led to the creation of several companies that deliver deepfakes as a product or even online service. Recorded Future has reported a threat actor's willingness to pay USD 16 000 for this kind of service.[29]

Since deepfakes are based on advanced AI and machine learning technologies, a high level of expertise is required to put the technology together. Accordingly, there are not as many threat actors with the skillset to develop them on their own as there are

25 University of Lincoln ScienceDaily, 'Two fraudsters, one passport: Computers more accurate than humans at detecting fraudulent identity photos,' 2019, accessed on 20 July, 2020, at www.sciencedaily.com/releases/2019/08/190801104038.htm.

26 Naser Damer, PhD. (n.d.). Fraunhofer IGD, 'Face morphing: a new threat?' accessed on 20 July 2020, at https://www.igd.fraunhofer.de/en/press/annual-reports/2018/face-morphing-a-new-threat.

27 David J. Robertson, et al. 'Detecting morphed passport photos: a training and individual differences approach,' Springer Nature, 2018, accessed on 20 July 2020, at https://cognitiveresearchjournal.springeropen.com/articles/10.1186/s41235-018-0113-8 .

28 Robin S.S. Kramer, et al., 'Face morphing attacks: Investigating detection with humans and Computers, Springer Nature, 2019, accessed on 20 July 2020, at https://link.springer.com/article/10.1186/s41235-019-0181-4.

29 Biometric update, 'Dark news from dark web: deepfakers are getting their act together', 2021, accessed on 16 March 2022, https://www.biometricupdate.com/202105/dark-news-from-dark-web-deepfakers-are-getting-their-act-together.

who would be interested in deepfakes as a service. Those who know how to leverage sophisticated AI can perform the service for others, enabling threat actors to manipulate a person's face and/or voice without understanding the intricacies behind how it works. Then they can conduct advanced social engineering attacks on unsuspecting victims, with the aim to make a sizable profit. Platforms offering these kinds of services have already started to emerge.[30]

Deepfake technology's impact on law enforcement

Law enforcement agencies will be adversely impacted by the rise of synthetic media and deepfakes. While they may provide some opportunities to benefit society, this report focuses on the malicious use of deepfakes. Adverse effects not only include the criminal uses described in the previous chapter, but also the more general impact of deepfakes on society. During foresight activities conducted by Europol, participants discussed how certain technologies could impact law enforcement. In relation to deepfakes, law enforcement agencies may even be forced into action, possibly the wrong action, by misinformation.

Impact on police work

Altered material on social media about events such as demonstrations may lead to police coming into action where it is not necessary, or in the wrong place. In police investigations, law enforcement may chase the wrong suspect of a crime when a deepfake version of the suspect fleeing a crime scene goes viral on social media, thereby giving the suspect the opportunity to get away.

Using deepfakes, people could falsely portray police officers committing transgressions in order to discredit the police or even incite violence against officers. In a time where distrust in authorities is growing, deepfakes and manipulated footage may be used to negatively affect public opinion. The impact of such images and footage is not to be underestimated, especially when this is combined with doxxing (exposing the identity of) the officers supposedly involved.

Impact on the legal process

In court, audio-visual evidence is usually trusted to be an authentic representation of events. Whether the file is extracted from the phone of a suspect, downloaded from social media, or received from the CCTV system of a shop near the crime scene, the authenticity of the scene depicted is not usually questioned. With the rise of deepfakes, it will become increasingly important

30 Europol, 'Malicious Uses and Abuses of Artificial Intelligence', 2020, accessed on 10 March 2022, https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence.

to scrutinise such content and verify if it is real or somehow artificially manipulated or generated.

Cross-checking footage will become even more important. It calls for a thorough vetting of digital evidence with specific attention to show it can be trusted to be authentic. A consistent and transparent chain of custody of digital evidence to prove no one could have doctored the evidence in the investigation is essential. For instance, as part of a child custody case, the mother of a child tried to convince the court that her husband behaved violently. She manipulated an audio recording of the man to make it look like he was making threats. Although this was not a real deepfake, it raises questions and concerns.[31] What if the manipulated footage remained unproven as fake?

With lighter-weight neural network structures and advances in hardware, training and generating time will be significantly reduced. In the near future, deepfake software will likely be able to generate full body deepfakes, real-time impersonations, and the seamless removal of elements within videos. The most recent algorithms can deliver increasingly higher levels of realism and run in near real time.

New capacities needed

Claims as to the use of deepfake material will require further law enforcement assessment, leading to new cases and new types of work. This will result in an increased workload and a push for law enforcement officers to develop new skills. Fake evidence has always existed and law enforcement agencies have procedures in place to assess the value of evidence. These procedures are developed for the types of forgeries already known and will have to be updated continuously with the rise of deepfakes. Law enforcement agencies will need to not only upskill their workforce to detect deepfakes, but also invest in their technical capabilities in order to address the upcoming challenges effectively while respecting fundamental rights.

Law enforcement agencies must consider this issue from multiple perspectives, when creating, storing, protecting and analysing audio-visual material. Specifically, they should:

- make use of tested and proven methods when making audio-visual recordings, e.g. certify a certain set-up for use in court and;
- employ technical and organisational safeguards against tampering, in order to be able to prove the authenticity of the footage.

Looking beyond law enforcement, general prevention strategies may be considered to make it harder to use deepfake technology

31 European Parliamentary Research Service, 'Tackling deepfakes in European policy', 2021, accessed 15 March 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

on audio-visual material. For example, technical solutions could be implemented to make deepfakes easier to spot or to increase markers of authenticity. The Content Authenticity Initiative[32] is an example of efforts to provide a standard for content authenticity and provenance.

Participants of the Innovation Lab's foresight activities anticipated new forms of crime, together with the resulting challenges in terms of data collection, criminal attribution and the heightened anonymity of the perpetrators, such as in creating deepfakes for criminal purposes. Criminals are likely to adopt new modus operandi that LEAs will be unable to identify or counter. The failure to legislate for these technologies will further stymie the investigative abilities of LEAs.

Mitigating these risks requires greater research and funding. Law enforcement professionals will need to anticipate possible crime scenarios such as those discussed in this report, and build out their investigative abilities accordingly. Furthermore, they should work with relevant stakeholders to ensure that the appropriate legislation is in place. Greater awareness building and transparency vis-a-vis the public is also needed to ensure the roll-out of these technologies is not hamstrung by concerns over privacy and data protection.

Deepfake detection

Law enforcement has always had to deal with fake evidence and therefore is in a good position to adapt to the presence of deepfakes. In order to handle the material LEAs encounter appropriately, it is important to account for the possibility of synthetic content with malicious intent. Here we discuss some of the ways this synthetic content can be uncovered, and preventative measures that can be taken against this threat.

Manual detection

It is still possible for the vast majority of deepfake content to be manually detected by looking for inconsistencies. This is a labour intensive task, which can only be done for a very limited number of files, and requires appropriate training to be familiar with all the relevant signs. Moreover, this process is further complicated by the human predisposition to believe audio-visual content and work from a truth default perspective.[33] That introduces the possibility of mistakes, both with selecting the files that need to be inspected as well as the inspection itself.

32 Content Authenticy Initiative, accessed on 10 March 2022, https://contentauthenticity.org.

33 Levine, T.R., 'Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection' Journal of Language and Social Psychology, 2014, pp. 378-392., https://www.researchgate.net/publication/273593306_Truth-Default_Theory_TDT_A_Theory_of_Human_Deception_and_Deception_Detection.

The models generating deepfakes might produce believable images, but these may still contain imperfections upon closer examination. A few examples include:

- blurring around the edges of the face;
- lack of blinking;
- light reflection in the eyes;
- inconsistencies in the hair, vein patterns, scars etc.;
- inconsistencies in the background, in subject as well as focus, depth etc.[34]

Automated detection

Ideally, a system would scan any digital content and automatically report on its authenticity. Such a system will most likely never be perfect, but with increased sophistication of deepfake technology, a high degree of certainty from such a system could be worth more than the manual inspection. There have already been efforts to create this kind of software from organisations such as Facebook[35] and security firm McAfee.[36] Detection software will look for signs of manipulation and help the reviewer decide on the authenticity with an explainable AI report on these signs.

As deepfake creation tools need training data to know what a real person looks like, most deepfake detection models are trained using databases of deepfake images. The learned signs of manipulation are thus based on data of known deepfakes, making it difficult to know how successful it will be at detecting deepfakes generated by unknown or updated models. Moreover, a deepfake GAN can be updated to account for the signs detected by known detection models in order to force the results to avoid producing these signs and henceforth go undetected.

Some examples[37] of detection technologies that have been developed in recent years are:

Biological signals

This approach tries to detect deepfakes based on imperfections in the natural changes in skin colour that arise from the flow of blood through the face.[38]

34 Venema, A. E., & Geradts, Z. J., 'Digital Forensics Deepfakes and the Legal Process,' 2020, TheSciTechLawyer, 16(4), pp. 14-23.

35 Michigan State University, MSU, 'Facebook develop research model to fight deepfakes', 2021, accessed on 10 March 2022, https://msutoday.msu.edu/news/2021/deepfake-detection.

36 McAfee, 'The Deepfakes Lab: Detecting & Defending Against Deepfakes with Advanced AI', 2020, accessed on 10 March 2022, https://www.mcafee.com/blogs/enterprise/security-operations/the-deepfakes-lab-detecting-defending-against-deepfakes-with-advanced-ai.

37 AIM, 'Top AI-Based Tools & Techniques For Deepfake Detection', 2020, accessed on 24 September 2021, https://analyticsindiamag.com/top-ai-based-tools-techniques-for-deepfake-detection.

38 U. A. Ciftci, I. Demir and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2020.3009287.

Phoneme-viseme mismatches

For some words the dynamics of the mouth, viseme, are inconsistent with the pronunciation of a phoneme. Deepfake models may not correctly combine viseme and phoneme in these cases.[39]

Facial movements

This approach uses correlations between facial movements and head movements to extract a characteristic movement of an individual to distinguish between real and manipulated or impersonated content.[40]

Recurrent Convolutional Models

Videos consist of frames which are really just a set of images. This approach looks for inconsistencies between these frames with deep learning models.

However, there are also challenges facing deepfake detection technology.

- Detection algorithms are trained on specific datasets. A slight alteration of the method used to generate the deepfake may therefore prevent detection.
- An update to the discriminative model of a GAN for specific artefacts detected by these systems will fool the detection software.
- Videos may be compressed or reduced in size, which causes problems with the reduction in pixels and artefacts, making it harder to detect the inconsistencies the system looks for.
- It has been shown that databases may be manipulated to misclassify images with certain identifiers by adding an identifier to a small part of the dataset (e.g. applying a trigger to 5% of the images resulted in the misclassification of fake images with the trigger as real).[41]
- Increased image forensics and deepfake detection capabilities drive the increased quality of deepfake videos. GANs can catch up relatively easily; by updating the discriminator to evade the detector, the learning capacity based on feedback loops of those GANs will work to produce a deepfake that can fool the detector.[42]

39 Agarwal, S. et al., 'Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches', 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, accessed on 10 March 2022, https://www.ohadf.com/papers/AgarwalFaridFriedAgrawala_CVPRW2020.pdf.

40 Agarwal, S. et al., 'Protecting world leaders against deep fakes', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 38-45, 2019, accessed on 10 March 2022, http://www.hao-li.com/publications/papers/cvpr2019workshopsPWLADF.pdf.

41 Cao, X. and Gong, N.Z., 'Understanding the Security of Deepfake Detection' ArXiv, 2021, accessed on 18 October 2021, https://arxiv.org/abs/2107.02045.

42 Wired, 'Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them', 2020, accessed on 15 March 2022, https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect.

Preventive measures

Organisations that rely on some kind of authorisation by face or voice biometrics should assess the authorisation process as a whole. Increasing the robustness of this process is currently considered as a better course of action than solely implementing specific deepfake detection systems. Common checks are:

- using audio-visual authorisation rather than just audio;
- demanding live video connection;
- requiring random complicated acts to be performed live in front of the camera, e.g. move hands across the face.

How are other actors responding to deepfakes?

In order to address the challenges posed by deepfake technology, it is important to look into what kind of action other actors, including the online platforms where most deepfakes can and might be shared, are addressing this threat. This is also influenced by the current legislative framework, which can ask for mandatory or voluntary measures. In this section, this report will show some examples of key online service providers and companies and their anti-deepfake measures. This chapter will then examine the EU regulatory framework in this area.

Technology companies

Early in 2020, Meta (formerly Facebook) announced a new policy banning deepfakes from their platforms.[43] Meta said it would remove AI-edited content that would likely mislead people, but made it clear that satire or parodies using the same technology would still be permissible on the platforms. In order for law enforcement to assess and address the impact of deepfakes on its work, it needs to be aware of the policies technology companies have put in place, as it is likely that potential evidence or malicious content will be shared via these platforms. How technology companies such as Twitter and Meta regulate deepfake technology will have an extensive impact on how people will engage with and react to deepfakes.

Examples of company policies:

- Meta (which owns Facebook and Instagram) aims to remove deepfakes, or otherwise edited media, where "manipulation isn't apparent and could mislead, particularly in the case of video content." [44]
- TikTok bans "Digital Forgeries (Synthetic Media or Manipulated Media) that mislead users by distorting the truth of events

43 Becoming Human: Artificial Intelligence Magazine, 'A Look at Deepfakes in 2020', 2020, accessed on 15 March 2022, https://becominghuman.ai/a-look-at-deepfakes-in-2020-13d3fe2b6ef7.

44 Meta, 'Manipulated media', accessed on 10 March 2022, https://transparency.fb.com/en-gb/policies/community-standards/manipulated-media/.

and cause significant harm to the subject of the video, other persons, or society." [45]

- Reddit "does not allow content that impersonates individuals or entities in a misleading or deceptive manner." This explicitly includes deepfakes "presented to mislead, or falsely attributed to an individual or entity." [46]
- Youtube has an existing ban for manipulated media under the spam, deceptive practices and scam policies of their community guidelines.[47]

Many of the policies use 'intent' as their barometer for deciding whether or not to remove a deepfake. However, defining 'intent' might prove challenging and highly subjective, since it is based on the assessment of individual actors. Nonetheless, it seems that online platforms could play a pivotal role in helping victims of deepfake technology to identify the perpetrator, but how this looks in practice remains to be seen. Moreover, technology providers also have responsibilities in safeguarding positive and legal use of their technologies and cooperating with law enforcement.

In addition to the policies, various technology companies are working on deepfake detection technologies. Developing detection technologies became a priority during the COVID-19 pandemic, and has gained new attention during the current conflict between Russia and Ukraine.

- Meta said it had developed an AI tool that detects deepfakes by reverse engineering a single AI-generated image to track its origin.[48]
- Google has released a large dataset of visual deepfakes that has been incorporated into the FaceForensics benchmark.[49]
- Microsoft has launched the Microsoft Video Authenticator, which can analyse a still photo or video to provide a percentage chance of whether the media has been artificially manipulated.[50]

45 TikTok, 'Community Guidelines', accessed on 10 March 2022, https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok.

46 Reddit, 'Updates to Our Policy Around Impersonation', 2020, accessed on 10 March 2022, https://www.reddit.com/r/redditsecurity/comments/emd7yx/updates_to_our_policy_around_impersonation.

47 Google Support, 'Misinformation policies', accessed on 10 March 2022, https://support.google.com/youtube/answer/10834785.

48 Politico, 'POLITICO AI: Decoded: Big Tech on the AI Act — AI inventors — Deepfakes', 2021, accessed on 10 March 2022, https://www.politico.eu/newsletter/ai-decoded/politico-ai-decoded-big-tech-on-the-ai-act-ai-inventors-deepfakes.

49 Google AI Blog, 'Contributing Data to Deepfake Detection Research', 2019, accessed on 10 March 2022, https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

50 Microsoft, 'New Steps to Combat Disinformation', 2020, accessed on 10 March 2022, https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator.

European Union

Regarding legal trends, participants of the foresight activities noted that at both the national and regional level, European law is struggling to keep pace with the evolution of technology and the changing definitions of crime. Participants flagged the need to establish new regulatory frameworks. These should be sensitive to contemporary law enforcement challenges (particularly in the digital realm), as well as to changing ethical norms. Some participants anticipated greater regulation of the digital sphere in the coming decade.

The COVID-19 crisis brought more discussion around regulation of disinformation and deepfake detection tools, but also an increased use of video conferencing tools with adjustable backgrounds and other filters bringing manipulated digital realities into our daily lives. The European Parliament report, 'Tackling Deepfakes in European Policy', explains this and shows that the regulatory landscape in the European Union related to deepfakes "comprises a complex web of constitutional norms, as well as hard and soft regulations on both the EU and the Member State level".[51]

The most relevant regulatory framework for law enforcement in the area of deepfakes will be the AI regulatory framework – which is still at proposal level and not applicable yet - proposed by the European Commission. The framework takes a risk-based approach to the regulation of AI and its applications. Deepfakes are explicitly covered by the passage about "AI systems used to generate or manipulate image, audio or video content", and have to adhere to certain minimum requirements. Minimum requirements include marking content as deepfake to make clear that users are dealing with manipulated footage." [52]

Deepfake detection software used by law enforcement authorities falls in the category of 'high-risk', as it is considered to pose a threat to the rights and freedoms of individuals. Detection software used by law enforcement under the AI regulatory framework would only be permitted under strict safeguards, such as the employment of risk-management systems and appropriate data governance and management practices.[53]

51 European Parliament Research Service, 'Tackling deepfakes in European policy', 2021, accessed on 10 March 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

52 European Parliament Research Service, 'Tackling deepfakes in European policy', 2021, accessed on 10 March 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

53 European Parliament Research Service, 'Tackling deepfakes in European policy', 2021, accessed on 10 March 2022, https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf.

Conclusion

As this report shows, in order to effectively address the threats posed by deepfake technology, legislation and regulation need to take into account law enforcement needs. Within the regulatory framework, law enforcement, online service providers and other organisations need to develop their policies and invest in detection as well as prevention technology. Policymakers and law enforcement agencies need to evaluate their current policies and practices, and adapt them to be prepared for the new reality of deepfakes.

The strategic foresight activities conducted by the Europol Innovation Lab identified a series of challenges that LEAs will have to contend with in the decade ahead. In particular, they identified risks associated with digital transformation, the adoption and deployment of new technologies, the abuse of emerging technology by criminals, accommodating new ways of working and maintaining trust in the face of an increase of disinformation.

In the months and years ahead, it is highly likely that threat actors will make increasing use of deepfake technology to facilitate various criminal acts and conduct disinformation campaigns to influence or distort public opinion. Advances in machine learning and artificial intelligence will continue enhancing the capabilities of the software used to create deepfakes. According to experts, GANs, availability of public datasets and increased computing power will be the main drivers of deepfake development in the future and make them more difficult to distinguish from authentic content.

The increase in use of deepfakes will require legislation to set guidelines and enforce compliance. Additionally, social networks and other online service providers should play a greater role in identifying and removing deepfake content from their platforms. As the public becomes more educated on deepfakes, there will be increasing concern worldwide about their impact on individuals, communities, and democracies.

In the EU there are various policies and regulatory attempts to address deepfakes. However, law enforcement's use of technology to detect deepfakes is considered as 'high-risk', according to some proposals. Therefore, it will be very important to clarify which practices should be prohibited under the AI regulatory framework. In order to address the challenges faced with deepfakes, law enforcement agencies need to prepare and train for deepfake detection and ensure e-evidence integrity, developing their capacities as described in this report. The regulatory framework should also support law enforcement preparedness efforts.

The Europol Innovation Lab is continuously monitoring the development of disruptive technologies such as deepfakes.

About the Europol Innovation Lab

Technology has a major impact on the nature of crime. Criminals quickly integrate new technologies into their modus operandi, or build brand-new business models around them. At the same time, emerging technologies create opportunities for law enforcement to counter these new criminal threats. Thanks to technological innovation, law enforcement authorities can now access an increased number of suitable tools to fight crime. When exploring these new tools, respect for fundamental rights must remain a key consideration.

In October 2019, the Ministers of the Justice and Home Affairs Council called for the creation of an Innovation Lab within Europol, which would develop a centralised capability for strategic foresight on disruptive technologies to inform EU policing strategies.

Strategic foresight and scenario methods offer a way to understand and prepare for the potential impact of new technologies on law enforcement. The Europol Innovation Lab's Observatory function monitors technological developments that are relevant for law enforcement and reports on the risks, threats and opportunities of these emerging technologies. To date, the Europol Innovation Lab has organised three strategic foresight activities with EU Member State law enforcement agencies and other experts.

www.europol.europa.eu