European Parliament

Tackling deepfakes in European policy

STUDY

Panel for the Future of Science and Technology

EPRS | European Parliamentary Research Service

Scientific Foresight Unit (STOA)

PE 690.039 – July 2021

EN

Tackling deepfakes in European policy

The emergence of a new generation of digitally manipulated media capable of generating highly realistic videos – also known as deepfakes – has generated substantial concerns about possible misuse. In response to these concerns, this report assesses the technical, societal and regulatory aspects of deepfakes.

The assessment of the underlying technologies for deepfake videos, audio and text synthesis shows that they are developing rapidly, and are becoming cheaper and more accessible by the day. The rapid development and spread of deepfakes is taking place within the wider context of a changing media system.

An assessment of the risks associated with deepfakes shows that they can be psychological, financial and societal in nature, and their impacts can range from the individual to the societal level.

The report identifies five dimensions of the deepfake lifecycle that policy-makers could take into account to prevent and address the adverse impacts of deepfakes. The legislative framework on artificial intelligence (AI) proposed by the European Commission presents an opportunity to mitigate some of these risks, although regulation should not focus on the technological dimension of deepfakes alone. The report includes policy options under each of the five dimensions, which could be incorporated into the AI legislative framework, the proposed European Union digital services act package and beyond. A combination of measures will likely be necessary to limit the risks of deepfakes, while harnessing their potential.

AUTHORS

This study has been written by Mariëtte van Huijstee, Pieter van Boheemen and Djurre Das (Rathenau Institute, The Netherlands), Linda Nierling and Jutta Jahnel (Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Germany), Murat Karaboga (Fraunhofer Institute for Systems and Innovation Research, Germany) and Martin Fatun (Technology Centre of the Academy of Sciences of the Czech Republic - TC ASCR), with the assistance of Linda Kool (Rathenau Institute) and Joost Gerritsen (Legal Beetle), at the request of the Panel for the Future of Science and Technology (STOA) and managed by the Scientific Foresight Unit, within the Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliament.

ADMINISTRATOR RESPONSIBLE

Philip Boucher, Scientific Foresight Unit (STOA)

To contact the administrator, please e-mail stoa@ep.europa.eu

LINGUISTIC VERSION

Original: EN

Manuscript completed in July 2021.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

http://www.europarl.europa.eu/stoa (STOA website)
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)

PE 690.039
ISBN: 978-92-846-8400-7
doi: 10.2861/325063
QA-02-21-930-EN-N

Executive summary

1. Introduction

The emergence of a new generation of digitally manipulated media has given rise to considerable worries about possible misuse. Advancements in artificial intelligence (AI) have enabled the production of highly realistic fake videos, that depict a person saying or doing something they have never said or done. The popular and catch-all term that is often used for these fabrications is 'deepfake', a blend of the words 'deep learning' and 'fake'. The underlying technology is also used to forge audio, images and texts, raising similar concerns.

Recognising the technological and societal context in which deepfakes develop, and responding to the opportunity provided by the regulatory framework around AI that was proposed by the European Commission, this report aims at informing the upcoming policy debate.

The following research questions are addressed:

1. What is the current state of the art and five-year development potential of deepfake techniques? (Chapter 3)
2. What does the societal context in which these techniques arise look like? (Chapter 4)
3. What are the benefits, risks and impacts associated with deepfakes? (Chapter 5)
4. What does the current regulatory landscape related to deepfakes look like? (Chapter 6)
5. What are the remaining regulatory gaps? (Chapter 7)
6. What policy options could address these gaps? (Chapter 8)

The findings are based on a review of scientific and grey literature, and relevant policies, combined with nine expert interviews, and an expert review of the policy options.

2. Deepfake and synthetic media technologies

In this report, **deepfakes** are defined as **manipulated or synthetic audio or visual media that seem authentic, and which feature people that appear to say or do something they have never said or done, produced using artificial intelligence techniques, including machine learning and deep learning**.

Deepfakes can best be understood as a subset of a broader category of AI-generated 'synthetic media', which not only includes video and audio, but also photos and text. This report focuses on a limited number of synthetic media that are powered by AI: deepfake videos, voice cloning and text synthesis. It also includes a brief discussion on 3D animation technologies, since these yield very similar results and are increasingly used in conjunction with AI approaches.

Deepfake video technology

Three recent developments caused a breakthrough in image manipulation capabilities. First, computer vision scientists developed algorithms that can automatically map facial landmarks in images, such as the position of eyebrows and nose, leading to facial recognition techniques. Second, the rise of the internet – especially video- and photo-sharing platforms – made large quantities of audio-visual data available. The third crucial development is the increase in image forensics capacities, enabling automatic detection of forgeries. These developments created the pre-conditions for AI technologies to flourish. The power of AI lies in its learning cycle approach. It detects patterns in large datasets and produces similar products. It is also able to learn from the outputs of forensics algorithms, since these teach the AI algorithms what to improve upon in the next production cycle.

Two specific AI approaches are commonly found in deepfake programmes: Generative Adversarial Networks (GANs) and Autoencoders. GANs are machine learning algorithms that can analyse a set of images and create new images with a comparable level of quality. Autoencoders can extract information about facial features from images and utilise this information to construct images with a different expression (see Annex 3 for further information).

Voice cloning technology

Voice cloning technology enables computers to create an imitation of a human voice. Voice cloning technologies are also known as audio-graphic deepfakes, speech synthesis or voice conversion/swapping. AI voice cloning software methods can generate synthetic speech that is remarkably similar to a targeted human voice. Text-to-Speech (TTS) technology has become a standard feature of everyday consumer electronics, such as Google Home, Apple Siri and Amazon Alexa and navigation systems.

The barriers to creating voice clones are diminishing as a result of a variety of easily accessible AI applications. These systems are capable of imitating the sound of a person's voice, and can 'pronounce' a text input. The quality of voice clones has recently improved rapidly, mainly due to the invention of GANs (see Annex 3).

Thus, the use of AI technology gives a new dimension to voice clone credibility and the speed at which a credible clone can be created. However, it is not just the sound of a voice that makes it convincing. The content of the audio clip also has to match the style and vocabulary of the target. Voice cloning technology is therefore connected to text synthesis technology, which can be used to automatically generate content that resembles the target's style.

Text synthesis technology

Text synthesis technology is used in the context of deepfakes to generate texts that imitate the unique speaking style of a target. The technologies lean heavily on natural language processing (NLP). A scientific discipline at the intersection of computer science and linguistics, NLP's primary application is to improve textual and verbal interactions between humans and computers.

Such NLP systems can analyse large amounts of text, including transcripts of audio clips of a particular target. This results in a system that is capable of interpreting speech to some extent, including the words, as well as a level of understanding of the emotional subtleties and intentions expressed. This can result in a model of a person's speaking style, which can, in turn, be used to synthesise novel speech.

Detection and prevention

There are two distinct approaches to deepfake detection: manual and automatic detection. Manual detection requires a skilled person to inspect the video material and look for inconsistencies or cues that might indicate forgery. A manual approach could be feasible when dealing with low quantities of suspected materials, but is not compatible with the scale at which audio-visual materials are used in modern society.

Automatic detection software can be based on a (combination of) detectable giveaways, some of which are AI-based themselves:

- Speaker recognition
- Voice liveness detection
- Facial recognition
- Facial feature analysis
- Temporal inconsistencies
- Visual artefacts
- Lack of authentic indicators

The multitude of detection methods might look reassuring, but there are several important cautions that need to be kept in mind. One caution is that the performance of detection algorithms is often measured by benchmarking it against a common data set with known deepfake videos. However, studies into detection evasion show that even simple modifications in deepfake production techniques can already drastically reduce the reliability of a detector.

Another problem detectors face is that audio-graphic material is often compressed or reduced in size when shared on online platforms such as social media and chat apps. The reduction in the number of pixels and artefacts that sound and image compression create can interfere with the ability to detect deepfakes.

Several technical strategies may prevent an image or audio clip from being used as an input for creating deepfakes, or limit its potential impact. Prevention strategies include adversarial attacks on deepfake algorithms, strengthening the markers of authenticity of audio-visual materials, and technical aids for people to more easily spot deepfakes.

3. Societal context

Media manipulation and doctored imagery are by no means new phenomena. In that sense, deepfakes can be seen as just a new technological expression of a much older phenomenon. However, that perspective would fall short when it comes to understanding its potential societal impact. A number of connected societal developments help create a welcoming environment for deepfakes: the changing media landscape by means of online sharing platforms; the growing importance of visual communication; and the growing spread of disinformation. Deepfakes find fertile ground in both traditional and new media because of their often sensational nature. Furthermore, popular visual-first social media platforms such as Instagram, TikTok and SnapChat already include manipulation options such as face filters and video editing tools, further normalising the manipulation of images and videos. Concerningly, non-consensual pornographic deepfakes seem to almost exclusively target women, indicating that the risks of deepfakes have an important gender dimension.

Deepfakes and disinformation

Deepfakes can be considered in the wider context of digital disinformation and changes in journalism. Here, deepfakes are only the tip of the iceberg, shaping current developments in the field of news and media. These comprise phenomena and developments including fake news, the manipulation of social media channels by trolls or social bots, or even public distrust of scientific evidence.

Deepfakes enable different forms of misleading information. First, deepfakes can take the form of convincing misinformation; fiction may become indistinguishable from fact to an ordinary citizen. Second, disinformation – misleading information created or distributed with the intention to cause harm – may be complemented with deepfake materials to increase its misleading potential. Third, deepfakes can be used in combination with political micro-targeting techniques. Such targeted deepfakes can be especially impactful. Micro-targeting is an advertising method that allows producers to send customised deepfakes that strongly resonate with a specific audience.

Perhaps the most worrying societal trend that is fed by the rise of disinformation and deepfakes is the perceived erosion of trust in news and information, confusion of facts and opinions, and even 'truth' itself. A recent empirical study has indeed shown that the mere existence of deepfakes feeds distrust in any kind of information, whether true or false.

4. Benefits, risks and impacts

Deepfake technologies can be used for a wide variety of purposes, with both positive and negative impacts. Beneficial applications of deepfakes can be conceived in the following areas: audio-graphic productions; human-machine interactions (improving digital experiences); video conferencing; satire; personal or artistic creative expression; and medical (research) applications (e.g. face reconstruction or voice creation).

Deepfake technologies may also have a malicious, deceitful and even destructive potential at an individual, organisational and societal level. The broad range of possible risks can be differentiated into three categories of harm: psychological, financial and societal. Since deepfakes target individual persons, there are firstly direct psychological consequences for the target. Secondly, it is also clear that deepfakes can be created and distributed with the intent to cause a wide range of financial harms. Thirdly, there are grave concerns about the overarching societal consequences of the technology. An overview of the risks identified in this research are presented in the table below.

Overview of different categories of risks associated with deepfakes

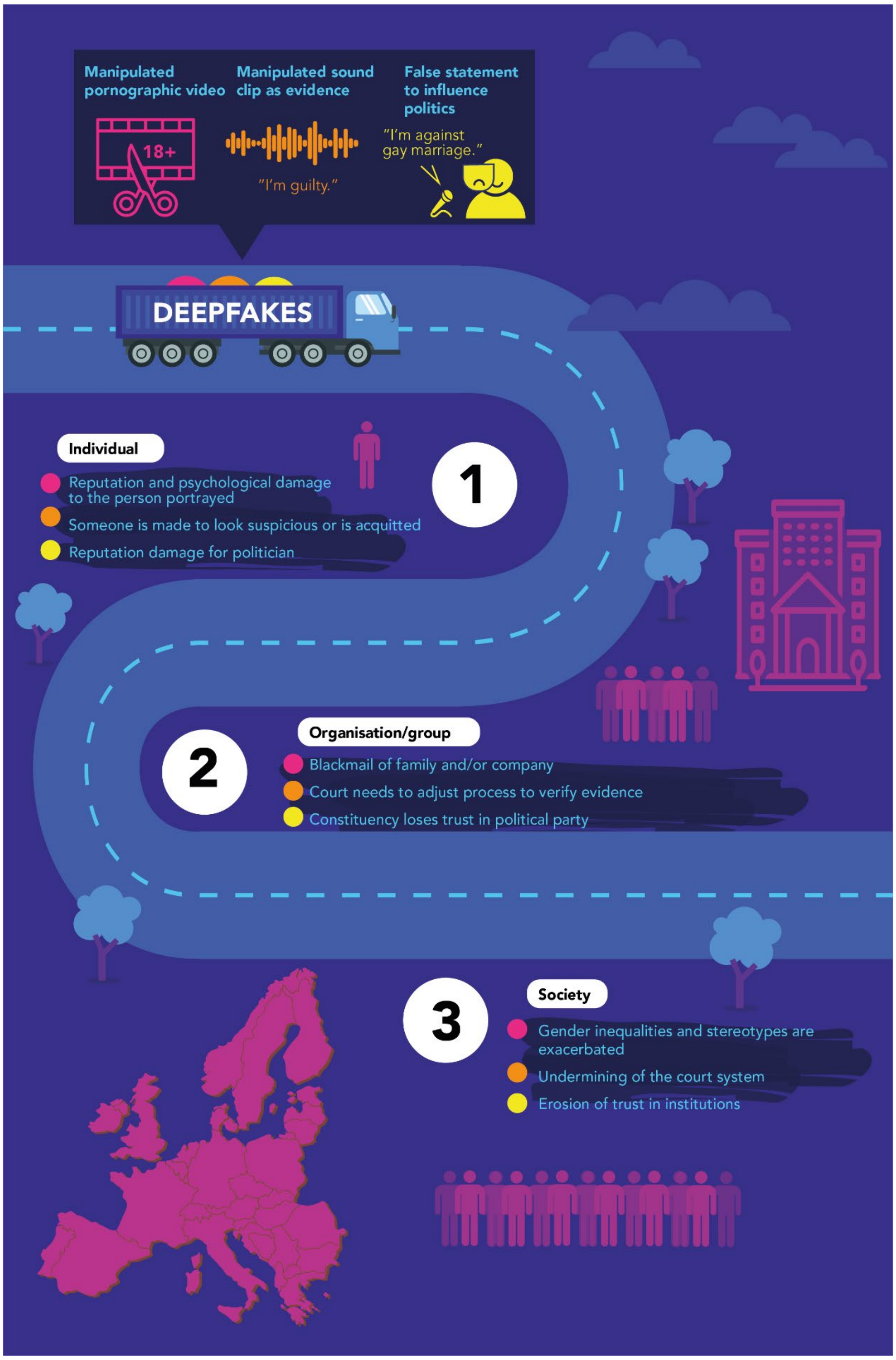
Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

5. Cascading impacts

The impact of a single deepfake is not limited to a single type or category of risk, but rather to a combination of cascading impacts at different levels (see infographic below). First, as deepfakes target individuals, the impact often starts at the individual level. Second, this may cause harm to a specific group or organisation. Third, the notion of the existence of deepfakes, a well-targeted deepfake, or the cumulative effect of deepfakes, may lead to severe harms on the societal level.

The infographic on the next page depicts three scenarios that illustrate the potential impacts of three types of deepfakes on the individual, group and societal levels: a manipulated pornographic video; a manipulated sound clip given as evidence; and a false statement to influence the political process.

Cascading effects of three types of deepfakes (a manipulated pornographic video, manipulated audio evidence and a false political statement) on the individual, organisational and societal level.


Manipulated pornographic video
18+
Manipulated sound clip as evidence
"I'm guilty."
False statement to influence politics
"I'm against gay marriage."
DEEPFAKES
Individual
1
Reputation and psychological damage to the person portrayed
Someone is made to look suspicious or is acquitted
Reputation damage for politician
Organisation/group
2
Blackmail of family and/or company
Court needs to adjust process to verify evidence
Constituency loses trust in political party
Society
3
Gender inequalities and stereotypes are exacerbated
Undermining of the court system
Erosion of trust in institutions

Source: Image created by Rathenau Instituut

6. Regulatory landscape and gaps

The regulatory landscape 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. On the European level, the most relevant policy trajectories and regulatory frameworks are:

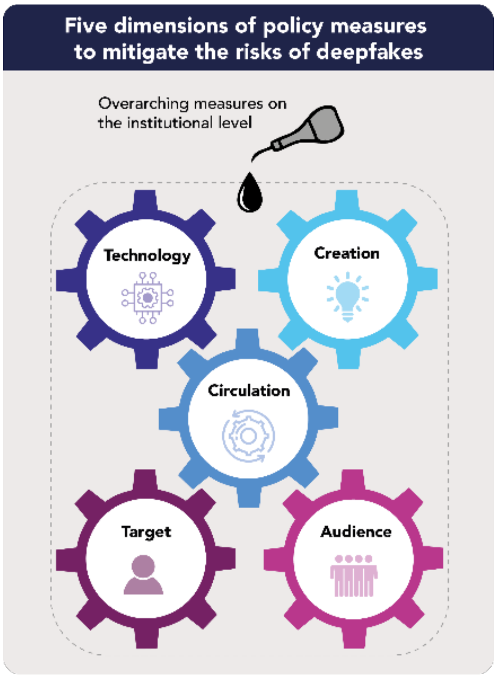
- The AI regulatory framework
- The General Data Protection Regulation
- Copyright regime
- e-Commerce Directive
- Digital services act
- Audio Visual Media Directive
- Code of Practice on Disinformation
- Action plan on disinformation
- Democracy action plan

Even though the current rules and regulations offer at least some guidance for mitigating potential negative impacts of deepfakes, the legal route for victims remains challenging. Typically, different actors are involved in the lifecycle of a deepfake. These actors might have competing rights and obligations. The scenarios in Chapter 7 illustrate how perpetrators often act anonymously, making it harder to hold them accountable. It seems that platforms could play a pivotal role in helping the victim to identify the perpetrator. Moreover, technology providers also have responsibilities in safeguarding positive and legal use of their technologies. This leads to the conclusion that policy-makers, when aiming to mitigate the potential negative impacts of deepfakes, should take different dimensions of the deepfake lifecycle into account.

7. Policy options

The report identifies various policy options for mitigating the negative impacts associated with deepfakes. In line with the different phases of the 'deepfake lifecycle', we distinguish five dimensions of policy measures: 1. the technology dimension, 2. the creation dimension, 3. the circulation dimension, 4. the target dimension, and 5. the audience dimension.

Five dimensions of policy measures to mitigate the risks of deepfakes


Five dimensions of policy measures to mitigate the risks of deepfakes
Overarching measures on the institutional level
Technology
Creation
Circulation
Target
Audience

Source: Rathenau Instituut

Technology dimension

The technology dimension covers policy options aimed at addressing the technology underlying deepfakes – AI-based machine learning techniques – and the actors involved in producing and providing this technology. The regulation of such technology lies largely within the domain of the AI regulatory framework as proposed by the European Commission. The framework takes a risk-based approach to the regulation of AI. Deepfakes are explicitly covered in the Commission proposal as 'AI systems used to generate or manipulate image, audio or video content', which have to adhere to certain minimum requirements, most notably when it comes to labelling. They are not included in the 'high risk' category, and uncertainty remains whether they could fall under the 'prohibited' category. The current AI framework proposal thus leaves room for interpretation. Since this research has documented a wide range of applications of deepfake technology, some of which are clearly high-risk, clarifications and additions to the AI framework proposal are recommended. Options include clarification of which AI practices should be prohibited under the AI framework; creation of legal obligations for deepfake technology providers; and regulation of deepfake technology as high-risk (for a full overview of the policy options identified, see Table 3).

Creation dimension

This dimension covers the policy options aimed at addressing the creators of deepfakes, or in AI framework terminology: the 'users' of AI systems. The AI framework proposal already formulates some rules and restrictions for the use of deepfake technology, but additional measures are possible. Options include clarification of the guidelines for the manner of labelling; limiting the exceptions to the deepfake labelling requirement; and banning certain applications altogether.

This dimension also addresses those who use deepfake technology for malicious purposes: the 'perpetrator'. Malicious users of deepfake technology often hide behind anonymity and cannot be easily identified, thereby escaping accountability. These users cannot be expected to willingly comply with the labelling requirement as introduced in the AI framework proposal. Policy measures needed against malicious users of deepfake technology therefore may include extending current legal frameworks with regard to criminal offences, diplomatic actions and international agreements to refrain from the use of deepfakes by foreign states and their intelligence agencies (for a full overview of the policy options identified, see Table 3).

Circulation dimension

This domain covers the policy options aimed at addressing the circulation of deepfakes, by formulating possible rules and restrictions for the dissemination of (certain) deepfakes. Online platforms, media and communication services play a crucial role in the dissemination of deepfakes. The dissemination and circulation of a deepfake to a large extent determines the scale and the severity of its impact. Therefore, responsibilities and obligations for platforms and other intermediaries are often recommended. Policy options that address this dimension mainly fit within the domain of the proposed digital services act, and include obliging platforms and other intermediaries to have deepfake detection software in place; increasing transparency obligations with regard to deepfake detection systems, detection results, and labelling and take-down decisions; and slowing down the speed of circulation (for a full overview of the policy options identified, see Table 3).

Target dimension

Malicious deepfakes create impacts at the individual level, for the person(s) depicted in the deepfake. This research has demonstrated that the rights of victims may be protected in principle, but it often proves difficult to effect this. Therefore, we offer several options for improving the protection of the victims, including institutionalising support for victims of deepfakes; strengthening the capacity of data protection authorities to respond to the use of personal data for deepfakes; and developing a unified approach for the proper use of personality rights within the European Union (for a full overview of the policy options identified, see Table 3).

Audience dimension

Deepfake impacts transcend the individual level and can cascade to group or even societal levels. Whether this happens partly depends on the audience response: will they believe the deepfake, disseminate deepfakes further when they receive them, lose trust in institutions? The audience dimension is therefore the final crucial dimension for policy-makers to limit the risks and impacts of deepfakes. Options listed here include the labelling of trustworthy sources; and investing in media literacy and technological citizenship (for a full overview of the policy options identified, see Table 3).

8. Conclusions

This research has identified numerous malicious as well as beneficial applications of deepfake technologies. These applications do not strike an equal balance, as malicious applications pose serious risks to fundamental rights. Deepfake technologies can thus be considered dual-use and should be regulated as such.

The invention of deepfake technologies has severe consequences for the trustworthiness of all audio-graphic material. It gives rise to a wide range of potential societal and financial harms, including manipulation of democratic processes, and the financial, justice and scientific systems. Deepfakes enable all kinds of fraud, in particular those involving identity theft. Individuals – especially women – are at increased risk of defamation, intimidation and extortion, as deepfake technologies are currently predominantly used to swap the faces of victims with those of actresses in pornographic videos.

Taking an AI-based approach to mitigating the risks posed by deepfakes will not suffice for three reasons. First, other technologies can be used to create audio-graphic materials that are effectively similar to deepfakes. Most notably 3D animation techniques may create very realistic video footage.

Second, the potential harms of the technology are only partly the result of the deepfake videos or underlying technologies. Several mechanisms are at play that are equally essential. For example, for the manipulation of public opinion, deepfakes need not only to be produced, but also distributed. Frequently, the policies of media broadcasters and internet platform companies are instrumental to the impact of deepfakes.

Third, although deepfakes can be defined in a sociological sense, it may prove much more difficult to grasp the deepfake videos, as well as the underlying technologies, in legal terms. There is an inherent subjective aspect to the seeming authenticity of deepfakes. A video that may seem convincing to one audience, may not be so to another, as people often use contextual information or background knowledge to make a judgement about authenticity.

Similarly, it may be practically impossible to anticipate or assess whether a particular technology may or may not be used to create deepfakes. One has to bear in mind that the risks of deepfakes do not solely lie in the underlying technology, but largely depend on its use and application. Thus, in order to mitigate the risks posed by deepfakes, policy-makers could consider options that address the wider societal context, and go beyond regulation. In addition to the technological provider dimension, this research has identified four additional dimensions for policy-makers to consider: deepfake creation; circulation; target/victim; and audience.

The overall conclusion of this research is that the increased likelihood of deepfakes forces society to adopt a higher level of distrust towards all audio-graphic information. Audio-graphic evidence will need to be confronted with higher scepticism and have to meet higher standards. Individuals and institutions will need to develop new skills and procedures to construct a trustworthy image of reality, given that they will inevitably be confronted with deceptive information. Furthermore, deepfake technology is a fast-moving target. There are no quick fixes. Mitigating the risks of deepfakes thus requires continuous reflection and permanent learning on all governance levels. The European Union could play a leading role in this process.

Table of contents

List of figures

List of tables

Glossary

Term	Explanation
Algorithm	A procedure or formula for solving a mathematical problem, generally based on a series of specified instructions.
Artificial intelligence	Refers to the development of computer systems able to perform tasks that normally require human intelligence.[1] A basic AI technique is **rule based** decision-making. This method essentially involves programming a series of 'if this, then that' instructions. More advanced techniques include machine learning and deep learning.[2]
Augmented reality	Technology that collects, analyses and applies data, and uses this to place digital layers over the physical reality in order to create 'hybrid' worlds. These are simultaneously physical and virtual. The technology is used for popular apps such as Snapchat and Pokémon Go.
Autoencoders	Unsupervised learning technique that can extract information about facial features in images, and utilise this information to construct portraits with a different expression.
Avatar	An icon or figure representing a particular person in a video game, smartphone app or online platform.[3]
Blockchain	Technology whereby a network of computers maintains a ledger and the computers determine between them whether changes in the ledger account are permitted. A familiar application of this technology is the virtual currency Bitcoin.[4]
Computer vision	AI-based technology that enables computers to gain high-level understanding from digital images or videos, and to accurately identify, classify and label objects.[5]
Deepfake	Manipulated or synthetic audio or visual media that seem authentic, which feature (a) person(s) that appear(s) to say or do something they never said or did, produced using artificial intelligence or machine learning.
Deep learning	Specific form of machine learning based on neural networks – inspired by the biology of the human brain – which combines different layers of information.[6]

[1] 'Artificial Intelligence,' Oxford Reference, accessed March 22, 2021.

[2] Pieter van Boheemen et al., 'Cyber Resilience with New Technology - An Opportunity and a Necessity' Rathenau Instituut, 2020.

[3] 'Avatar,' Lexico Dictionaries, accessed March 22, 2021.

[4] van Boheemen et al., 'Cyber Resilience with New Technology - An Opportunity and a Necessity.'

[5] 'What Is Computer Vision?,' IBM, accessed March 22, 2021.

[6] van Boheemen et al.

Disinformation	The conscious, usually covert, dissemination of misleading information with the aim of causing damage to the public debate, democratic processes, the open economy or national security.[7]
Facial recognition	Computer vision-based method that can be used for identifying or authenticating the identity of a specific person from their facial landmarks. The technology relies heavily on comparison of training data.
Forensics	Algorithms that can be used to detect media manipulations and forgery. Image forensics in deepfakes, for example, can be used to detect the lack of eye-blinking or other inconsistencies.
GANs	Generative Adversarial Networks are machine learning algorithms that can analyse a given set of images and create new images with a similar level of quality.
Machine learning	Algorithms with a certain unsupervised learning capacity. Machine learning is more advanced than rule-based AI, and is generally based on the comparison of data rather than prior instructions. The technology relies heavily on statistics.[8]
Model	In computer science, a model aims to express the terms and concepts used by the domain experts to discuss a concept and find relationships between different concepts.
Natural language processing	Specific subfield of artificial intelligence that gives computers the ability to read, understand and derive meaning from human language, like speech and text. Ultimately, this technology can be used to produce and manipulate natural language.[9]
Sharing platforms	Digital platforms where information like video, photo or text can be shared. Sharing platforms play a pivotal role in the development and use of new technologies like augmented reality and deepfakes. Well-known examples include Facebook, TikTok, Snapchat and Twitter.
Synthesis	Umbrella term for artificially created media, including text, speech and image synthesis.
Synthetic media	All-encompassing term for different sorts of automatic and artificial media productions, including video, audio and text manipulation. Synthetic media are commonly based on artificially intelligent software.
Virtual Reality	Technology that is used for creating an immersive computer-generated three-dimensional environment. VR makes new digital experiences and forms of communication possible. Unlike augmented reality, it implies a complete immersion experience that shuts out the physical world.[10]

[7] 'Kamerbrief over Beleidsinzet Bescherming Democratie Tegen Desinformatie,' Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2019.

[8] van Boheemen et al., 'Cyber Resilience with New Technology - An Opportunity and a Necessity.'

[9] Kyrill Poelmans, 'What Is Natural Language Processing (NLP)?,' Textmetrics, June 25, 2020.

[10] Dhoya Snijders et al., 'Responsible VR. Protect Consumers in Virtual Reality' Rathenau Instituut, 2020.

List of abbreviations

AI	Artificial intelligence
CGI	Computer generated imagery
DLT	Distributed ledger technology
GANs	Generative adversarial networks
NLP	Natural language processing
TTS	Text to speech
3D	Three-dimensional

1. Introduction

1.1. Background

The emergence of a new generation of digitally manipulated media is currently drawing a lot of attention from researchers, policy-makers and the public at large. Advances in artificial intelligence (AI) have enabled the production of highly realistic fake videos, that depict a person saying or doing something they have never said or done.[11] The popular and catch-all term that is often used for these fabrications is 'deepfake'; a blend of the words 'deep learning' and 'fake'. Prominent examples of deepfake videos include former United States President Barack Obama offending his successor Donald Trump, and Donald Trump calling on the Belgian government to withdraw from the Paris Climate Agreement. However, videos like these might just be the tip of the iceberg.

Deepfakes can best be understood as a subset of a broader category of AI-generated 'synthetic media', which not only includes video and audio, but also photos and text. The rapid upswing of this type of media creation and manipulation has given rise to considerable concerns about possible misuse. Some commentators argue that the production and distribution of deepfakes by malicious actors has the potential to be used to extort, humiliate, harass, and blackmail victims,[12] and could leave individuals, companies, and government institutions vulnerable.[13]

However, the impact of deepfakes and synthetic media could be even larger than the direct impact of an actual abuse of these media.[14] Some foresee the possibility of an 'infocalypse', a grim vision of a future in which we can no longer believe digital content because anything we see might well be manipulated.[15] Such an outcome could cause an erosion of trust and would have a destabilising impact on society as a whole. As such, deepfakes not only pose a challenge to privacy, but also to democracy and national security,[16] as they could undermine trust in public discourse when used as a vehicle for misinformation.[17]

Media manipulation is anything but a new phenomenon. So why have synthetic media, and deepfakes in particular, recently spurred so much interest? Firstly, the current technological advances are quickly driving the improvement of the quality of deepfakes and makes it harder to distinguish fake from real.[18] The visceral immediacy of audio-visual manipulations gives unprecedented impact and authority.[19] The audio-visual element of deepfakes has a more powerful effect on our psychology than other types of media.[20] Moreover, the technology that is used to deploy deepfakes is getting cheaper, more

[11] Jon Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario' Carnegie Endowment for International Peace, 2020.

[12] Douglas Harris, 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You,' *Duke Law & Technology Review* 17, no. 1 (January 5, 2019): 99–127.

[13] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.'

[14] Matteo Bonfanti, 'The Weaponisation of Synthetic Media: What Threat Does This Pose to National Security?,' *CSS ETH Zurich* (blog), 2020.

[15] Nina Schick, *Deep Fakes and the Infocalypse* Octopus Publishing Group, 2020.

[16] Robert Chesney and Danielle Keats Citron, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,' SSRN Scholarly Paper Rochester, NY: Social Science Research Network, July 14, 2018.

[17] Cristian Vaccari and Andrew Chadwick, 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,' *Social Media + Society* 6, no. 1 (January 1, 2020): 2056305120903408.

[18] Bonfanti, 'The Weaponisation of Synthetic Media.'

[19] Martijn Rasser, 'Why Are Deepfakes So Effective?,' Scientific American Blog Network, 2019.

[20] Carly Minsky, ''Deepfake' Videos: To Believe or Not Believe?,' January 26, 2021.

accessible and more user-friendly, meaning that creators no longer need to be very tech-savvy. This makes the prospect of mass production and distribution of deepfakes very likely.

Secondly, several societal factors and trends make the advent of deepfakes more salient. Citron and Chesney[21] argue that the proliferation of synthetic media 'comes at a perilous time', since news is no longer only distributed by trusted media companies. These changes in the media landscape evolving in the direction of user-generated content not only have consequences for the dissemination of deepfakes, but also makes it harder, if not impossible, to agree on ethical or professional codes of conduct, or norms and behaviours. Simultaneously, imagery, especially online, is becoming the dominant mode of expression.[22] As a result, deepfakes might become a prominent source of information power in the near future.

1.2. Research questions

Recognising the technological and societal context in which deepfakes develop, and responding to the opportunity provided by the regulatory framework around AI that is being developed by the European Commission, the authors of this report aim to inform the upcoming policy debate.

The following research questions are addressed:

1. What is the current state of the art and five-year development potential of deepfake techniques? (Chapter 3)
2. What does the societal context in which these techniques arise look like? (Chapter 4)
3. What are the benefits, risks and impacts associated with deepfakes? (Chapter 5)
4. What does the current regulatory landscape related to deepfakes look like? (Chapter 6)
5. What are the remaining regulatory gaps? (Chapter 7)
6. What policy options are possible to address these gaps? (Chapter 8)

1.3. Definitions

The term deepfake is mostly used to refer to AI-generated video-graphic media. Deepfakes are commonly seen as a specific branch of a broader spectrum of computer-generated content known as 'synthetic media'. The meaning of the word 'synthetic' in this term is similar to 'synthetic rubber'. It signals that the term encompasses imitations of text, audio-, photo- and video-graphic materials that are perceived as authentic. In popular media, the terms deepfake and synthetic media are seemingly interchangeable, for example, describing AI-generated voice as 'deepfake voice' or 'synthetic voice'. To prevent misunderstandings, this paragraph describes how we define deepfakes and synthetic media and how these definitions limit the scope of the report.

1.3.1. Deepfakes

In this report, 'deepfakes' are understood as **manipulated or synthetic audio or visual media that seem authentic, and which feature (a) person(s) that appear(s) to say or do something they have never said or done, produced using artificial intelligence techniques, including machine learning and deep learning**.

We limit the scope of the definition to media that feature a **human being**, although we are aware of the fact that the same technology is applied to videos containing other subjects.

[21] Chesney and Citron, 'Deep Fakes.'

[22] Ignas Kalpokas, 'Problematising Reality: The Promises and Perils of Synthetic Media,' *SN Social Sciences* 1, no. 1 (November 9, 2020): 1.

The term deepfake originates from Reddit, a popular internet message board. In 2017, an anonymous user called deepfake posted manipulated videos and shared the programming code that enabled others to follow suit.[23] The first videos contained pornographic content, in which the faces of the original actresses were replaced by those of celebrities Taylor Swift, Scarlett Johansson and Gal Gadot.[24] The term deepfake attracted much attention when these videos were reported in mainstream media.[25] The term was further established when non-pornographic videos were published, featuring well-known personas such as the actor Nicholas Cage[26], US President Barack Obama and Facebook Chief Executive Officer (CEO) Mark Zuckerberg.[27] The recent introduction of the term is also reflected in scientific literature. Only a few dozen papers mention the term in 2018, while several hundred papers covered the subject in 2020.[28]

The underlying technologies that enable the creation of deepfakes thus date back further than the term itself. The manipulation of videos is probably as old as video itself, since every step between the capturing of the light of a scene to displaying an image can be considered a manipulation. Recently, however, several technical and societal developments have led to the adoption of a new term. Most significant were a number of developments in computer vision and machine learning technology. In Chapter 3, the background to these developments will be further explained.

In the media, the term deepfake is used to refer to a wide spectrum of techniques that result in manipulated videos.[29] For example, the British Centre for Data Ethics and Innovation believes the term should be used regardless of the object in the video.[30] Many others require some element of impersonation.[31] Some academics discriminate between the degree or specific area of manipulation, such as manipulating facial expressions, swapping identities or completely generating novel faces.[32] A number of academics describe the use of **artificial intelligence** as a crucial hallmark of deepfakes.[33]

The experts consulted for this research noted that the term deepfake has a negative connotation or points to a malicious intent. This is probably due to the fact that the word 'fake' is often associated with unlawful acts, such as fraud and forgery. Nevertheless, this connotation apparently does not prevent some companies from offering legitimate products and services branded as deepfakes.

[23] The Reddit user most likely came up with the term deepfake by combining the terms 'Deep Learning' (a specific approach to Artificial Intelligence) and 'fake'. Although the term is clearly intended as a portmanteau, alternative stylings are in use, such as the hyphenated 'deep-fake' or two words writing 'deep fake'. See Ben Sasse, 'S.3805 - 115th Congress (2017-2018): Malicious Deep Fake Prohibition Act of 2018,' webpage, December 21, 2018.

See also 'Words We're Watching: 'Deepfake,'' Merriam Webster, accessed April 28, 2021.

[24] Rebecca Delfino, 'Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act,' *Fordham Law Review* 88, no. 3 (December 1, 2019): 887.

[25] Samantha Cole, 'We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now,' 2018.

[26] Sam Haysom, 'People Are Using Face-Swapping Tech to Add Nicolas Cage to Random Movies and What Is 2018,' Mashable, 2018.

[27] L. Whittaker et al., ''All around Me Are Synthetic Faces': The Mad World of AI-Generated Media,' *99,* 2020.

[28] Thanh Thi Nguyen et al., 'Deep Learning for Deepfakes Creation and Detection: A Survey,' *ArXiv:1909.11573 [Cs, Eess]*, July 28, 2020.

[29] Britt Paris and Joan Donovan, 'Deepfakes and Cheap Fakes,' Data & Society, September 18, 2019.

[30] 'Snapshot Paper - Deepfakes and Audiovisual Disinformation' Centre for Data Ethics and Innovation, 2019.

[31] Chesney and Citron, 'Deep Fakes.'

[32] Ruben Tolosana et al., 'DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,' *ArXiv:2001.00179 [Cs]*, June 18, 2020.

[33] Johannes G. Botha and Heloise Pieterse, *Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security*, 2020.

1.3.2. Synthetic media

In the early stages of the research for this report, it became clear that limiting the scope solely to what is strictly understood as deepfakes would not fully address the rising concerns regarding computerised manipulations. At first glance, the term 'synthetic media' seems more appropriate, as it encompasses all kinds of automatic and artificial media productions and manipulations.

However, widening the scope to the full breadth of synthetic media would result in such a broad field that the result would no longer be contained within the report's goal to offer policy options for legislative efforts on AI. The report therefore focuses on a number of AI-powered synthetic media only: deepfake videos, voice cloning and text synthesis. The only exception is a brief discussion on 3D animation technologies, since these yield very similar results and are increasingly used in conjunction with AI approaches.

1.4. Outline

- Chapter 2: Methodology
- Chapter 3: Deepfake and synthetic media technologies
- Chapter 4: Societal context
- Chapter 5: Benefits, risks and impacts of deepfakes
- Chapter 6: Regulatory landscape
- Chapter 7: Regulatory gaps
- Chapter 8: Regulatory options

2. Methodology

The findings reported in this document are the result of research based on literature review, expert interviews and expert review. This section describes how these methods were applied.

2.1. Literature review and analysis

This report is based primarily on a review of primary and secondary scientific literature. Searches were conducted in several literature databases, including Scopus, Google Scholar, IEEE Explore and SSRN. Given the recent emergence of the term 'deepfake' the search was limited to articles published in 2019, 2020 and 2021. Based on the titles and abstracts all articles were categorised according to the already defined structure of this report and subsequently analysed. Additional literature was collected based on hand-searching of references in the identified articles, taking special note of frequently cited articles. Other keywords were 'disinformation', 'image synthesis', 'neural language model', 'regulation AND artificial intelligence' and 'voice cloning'. Since the keyword 'voice cloning' yielded only a small number of results, additional keywords were added to the search, including 'speech synthesis', 'audio generation model' and 'text-to-speech'.

Additionally, the literature study was extended by searches for reports covering deepfakes by (international) commercial and non-profit private organisations as well as public institutions. This resulted in a list of 35 reports from organisations such as Brookings, Carnegie Endowment, Democracy Reporting International, Electronic Frontier Foundation, European Union Agency for Cybersecurity (ENISA), Europol, Mozilla, NATO, Sensity (formerly known as Deeptrace), World Economic Forum and Witness Media Lab. To distinguish between EU and international legislations and perspectives, all reports were indexed based on country of origin of the institution. Next, the reports were scanned for relevant content in relation to the structure of this report.

To get a sense of the practical use and current trends in deepfake content and technology, online communities were explored. During the course of the study, we conducted several searches on YouTube for 'deepfake', and we scanned online forums on Reddit, MrDeepfake, and the communities around popular deepfake software such as Deepfacelab and Faceswap.

For the European policy analysis in this report, a distinct number of existing and developing instruments of the European Union were analysed, including:

- AI legislative framework
- GDPR
- Copyright Law
- Image rights
- E-commerce Directive
- digital services act
- Audio Visual Media Directive
- Code of Practice on Disinformation
- Action plan against disinformation
- Democracy action plan

2.2. Expert interviews

The outcomes of the literature review and analysis were supplemented by expert interviews. Nine experts were identified in the literature based on their expertise with regard to the technology and main impact areas (Annex 1: List of experts). The interviews were conducted in a semi-structured fashion, based on a predefined list of questions (Annex 2: Interview questions). The insights from the expert interviews are integrated in the analysis throughout the report.

2.3. Expert review

The research team drafted a wide array of policy options based on the literature review and analysis combined with insights from the interviews. These policy options were then reviewed by three expert reviewers (Annex 1: Reviewers), which led to further refinement and improvement of the policy options in Chapter 8 and Table 3.

3. Deepfake and synthetic media technologies

This chapter describes the technological aspects of photo- and video-graphic deepfakes, audio-graphic deepfakes (voice cloning) and text synthesis.

3.1. Photo- and video-graphic deepfake technology

Photo- and video-graphic deepfakes are created by similar technologies. Videos are simply converted into photos by splitting every frame. Next, each image is manipulated separately.

Image manipulation technology gradually evolved over time

The methods and level of sophistication for such manipulations have gradually increased over the past decades. When computers were equipped with graphical user interfaces in the 1970s the first applications for image manipulation were developed as well. When Photoshop became popular in the 1990s a broad audience gained the ability to manipulate images.

High-quality video manipulation, however, was until recently primarily conducted by professionals from the cinematographic industry and academics in the field of image processing. Automatic manipulations that are similar to what we understand as deepfakes today already started to appear in the 1990s, such as the Video Rewrite Program that synthesised facial animations of US president John F Kennedy in 1997.[34] As computing power increased over time, movie studios developed Computer-Generated Imagery (CGI) technology and distributed the results in cinemas around the world. A well-known example is the winner of the 2009 Academy Award for Best Visual Effects: *The Curious Case of Benjamin Button*. Throughout the entire movie, computer-aided manipulations of the face of actor Brad Pitt are used to create the illusion of reverse ageing.

Recent breakthrough technological progress

Three recent developments caused a breakthrough in image manipulation capabilities. First, computer vision scientists developed algorithms that can automatically map facial landmarks in images such as the position of eyebrows and nose, leading to facial recognition techniques. Simultaneously, the rise of the internet – especially video- and photo-sharing platforms, such as YouTube – made large quantities of audio-visual data available. Today, data sets are widely available containing large quantities of pre-labelled images and videos of celebrities.[35] This also explains why celebrities and public figures such as US President Barack Obama were among the first to appear in deepfake videos. [36] The third crucial development is the increase in image forensics capacities, enabling automatic detection of forgeries.

The above-mentioned developments are three important pre-conditions for AI technologies to flourish. AI can gain from a learning approach when large data sets are available combined with the ability to gain feedback. Therefore, forensics algorithms are crucial.

Two specific AI approaches are commonly found in deepfake programmes: Generative Adversarial Networks (GANs) and Autoencoders. GANs are machine learning algorithms that can analyse a set of images and create new images with a comparable level of quality. Autoencoders can extract

[34] Christoph Bregler, Michelle Covelle, and Malcolm Slaney, 'Video Rewrite,' Interval Research Corporation, 1997.

[35] Yuezun Li et al., 'Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,' *ArXiv:1909.12962 [Cs, Eess]*, March 16, 2020.

[36] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, 'Synthesizing Obama: Learning Lip Sync from Audio,' *ACM Transactions on Graphics* 36, no. 4 (July 20, 2017): 95:1-95:13.

information about facial features in images, and utilise this information to construct images with a different expression. In Annex 3, we describe these techniques in more detail.

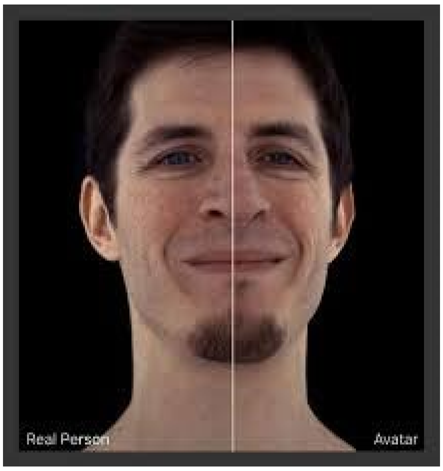
3D avatar animation technology

3D animation technology is increasingly able to generate videos with a similar quality to AI-based deepfake technology. Some deepfake programmes even combine AI image generation and 3D animation (see Paragraph on Trends). Most notably are avatar technologies that animate 3D models of a person's head or entire body.

These programmes first create a photorealistic 3D model either manually, or automatically by deriving the 3D landmarks from a single image or multiple images of a person. Next, the 3D model can be animated by capturing the movements from an actor, or by programmatically animating the model based on the interpretation of an audio-graphic speech fragment or text.

3D facial animation techniques were until recently mostly applied in cinema movies and computer games. In the past five years, the popularity of Virtual Reality and Augmented Reality technology has increased, due to the availability of equipment at a consumer-friendly price. Large technology companies, such as Facebook, are also investing in technological developments. Their desire to let users control a realistic virtual representation of themselves in a 3D environment has led to the development of products such as Facebook Codec Avatar. In demonstration videos, the company shows that it is difficult for an audience to tell the difference between a video of a real person and one that is generated using their 3D avatar technology (See Figure 1).[37]

Figure 1 - Facebook Codec Avatar



3.2. Specific graphical deepfake techniques

Within the realm of deepfake techniques, several specific applications can be discerned. The technologies described above can, for example, be applied to specific parts of an image or entire frames from a video, resulting in specific outcomes that are often described as discrete deepfake techniques. In the table below we list frequently used terms that refer to these specific techniques, accompanied by a brief description. Below the table, a collection of examples is presented.

[37] *Codec Avatars Side-by-Side Comparison* Tech@Facebook, 2019.

Table 1 - Five graphical deepfake techniques

Techniques	Description
Facial expression manipulation	These techniques can be used to modify a specific part of a target's face in an image or video, preserving the target's identity (See Figure 2). For example, this can be achieved by transferring the expressions of an actor to the target (facial re-enactment).[38] Similar techniques are used for 'visual dubbing' purposes in which only the movement of the lips of a target are adjusted based on the modification of audio or by using text input.[39] Any part of a person's face can be targeted, including adjusting the lighting or pose of the head.[40]
Face morphing	The goal of this technique is to create an image or video in which the faces of similar-looking people are merged in such a way that the pictures seem to depict both (See Figure 3). It is, for example, used to fraudulently obtain authentic identification documents, such as passports, that can be used by multiple persons.[41]
Face replacement/swap	With these techniques, the face of a target person is replaced by the face of the source video (See Figure 4).[42] Popular tools are Faceswap, DeepFaceLab and DFaker[43]. Alternatively, a face can be replaced with footage rendered based on a 3D model.[44]
Face generation	Face generators synthesise partial or entirely new images of people that do not exist (See Figure 5). The technique makes use of GANs.[45] Partial generators can, for example, be used to replace the VR goggles of a person by an image of their eyes.[46]
Full body puppetry	These techniques enable users to modify the pose of a part or entire body of a target in an image or video. An existing video could be used as a driver, or a sequence that was recorded using motion capture. This technology can, for example, make it appear as if anyone can dance like a professional[47].

[38] Justus Thies et al., 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos,' *Computer Vision Foundation*, 2016.

[39] Hyeongwoo Kim et al., 'Neural Style-Preserving Visual Dubbing,' *ACM Transactions on Graphics* 38, no. 6 (November 8, 2019): 1–13.

[40] Zhenliang He et al., 'AttGAN: Facial Attribute Editing by Only Changing What You Want,' *ArXiv:1711.10678 [Cs, Stat]*, July 25, 2018.

[41] Naser Damer et al., 'MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network,' in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018, 1–10.

[42] Dmitri Bitouk et al., 'Face Swapping: Automatically Replacing Faces in Photographs,' *ACM Transactions on Graphics* 27, no. 3 (August 1, 2008): 1–8; Yuval Nirkin et al., 'On Face Segmentation, Face Swapping, and Face Perception,' *ArXiv:1704.06729 [Cs]*, April 21, 2017; Yuval Nirkin, Yosi Keller, and Tal Hassner, 'FSGAN: Subject Agnostic Face Swapping and Reenactment,' *ArXiv:1908.05932 [Cs]*, August 16, 2019; Lingzhi Li et al., 'FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping,' *ArXiv:1912.13457 [Cs]*, September 15, 2020; Iryna Korshunova et al., 'Fast Face-Swap Using Convolutional Neural Networks,' *ArXiv:1611.09577 [Cs]*, July 27, 2017.

[43] Deepfake, 'Faceswap,' *Github Repository*, 2021; iperov, 'DeepFaceLab,' *Github Repository*, 2021; dfaker, 'Df,' *Github Repository*, 2021.

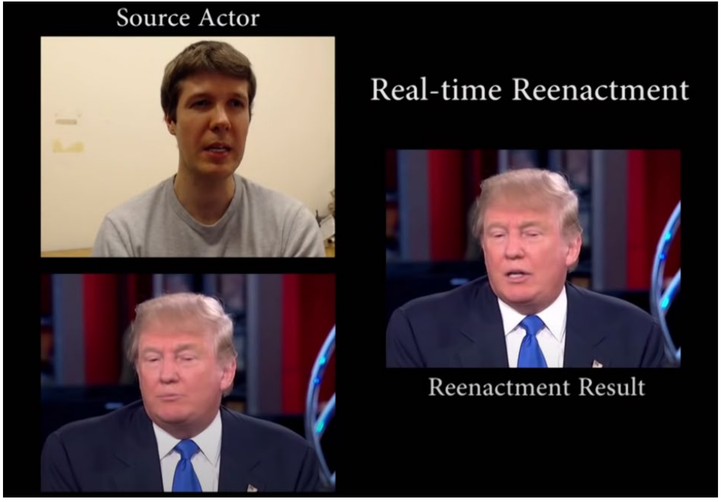
[44] Yi-Ting Cheng et al., '3D-Model-Based Face Replacement in Video,' in *SIGGRAPH '09: Posters*, SIGGRAPH '09 New Orleans, Louisiana: Association for Computing Machinery, 2009, 1.

[45] Yunjey Choi et al., 'StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation,' *ArXiv:1711.09020 [Cs]*, September 21, 2018; Tero Karras, Samuli Laine, and Timo Aila, 'A Style-Based Generator Architecture for Generative Adversarial Networks,' *ArXiv:1812.04948 [Cs, Stat]*, March 29, 2019; Tero Karras et al., 'Analyzing and Improving the Image Quality of StyleGAN,' *ArXiv:1912.04958 [Cs, Eess, Stat]*, March 23, 2020; Jianmin Bao et al., 'Towards Open-Set Identity Preserving Face Synthesis,' *ArXiv:1803.11182 [Cs]*, August 9, 2018.

[46] Matthias Niessner, *Face2Face: Real-Time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)*, 2016.

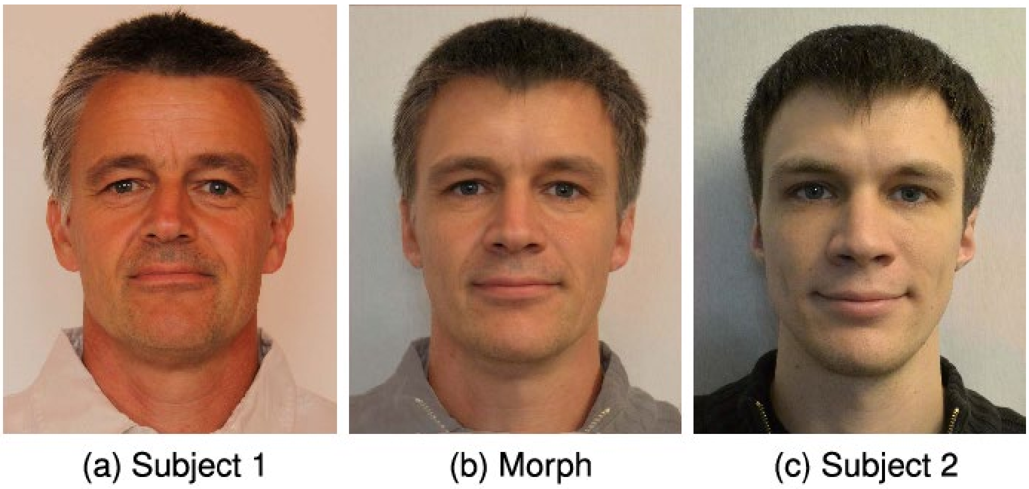
[47] Caroline Chan et al., 'Everybody Dance Now,' *ArXiv:1808.07371 [Cs]*, August 27, 2019.

Figure 2 - Facial expression manipulation[48]



Using real-time video input (top left) as a driver, the target image (bottom left) is transformed into a real-time video output (right).

Figure 3 - Face-morphing example[49]

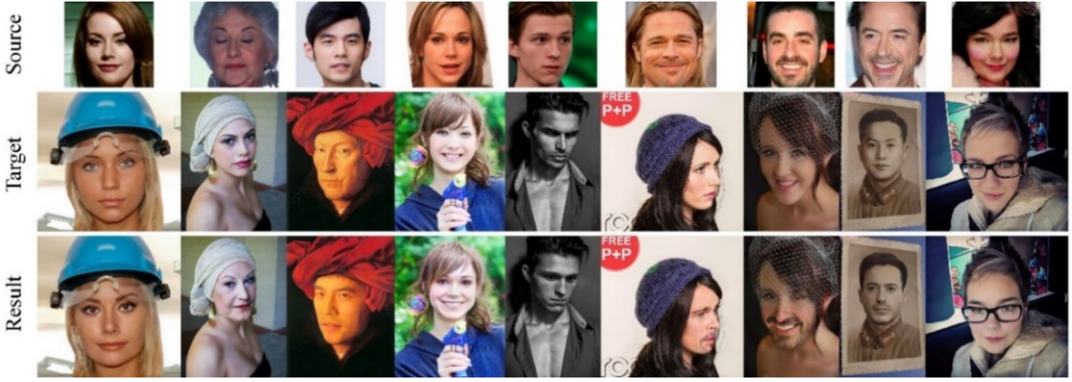


The pictures of the individuals on the left and right are morphed into a picture that seems to depict both (middle).

[48] Niessner, *Face2Face*.

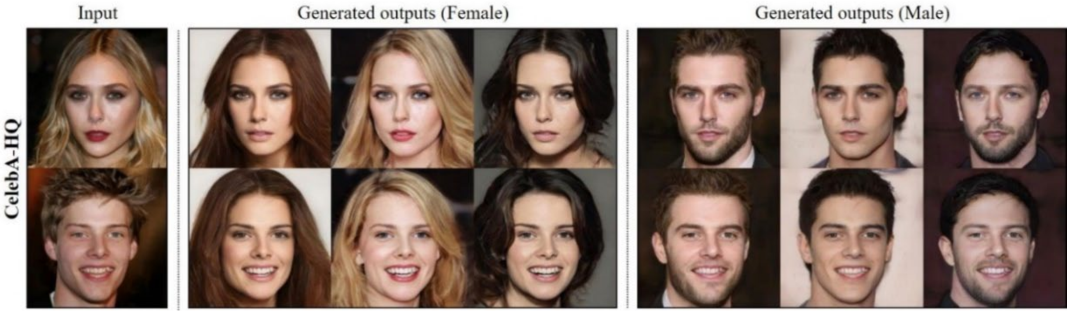
[49] Ulrich Scherhag et al., 'Detection of Face Morphing Attacks Based on PRNU Analysis,' *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.

Figure 4 - Face replacement



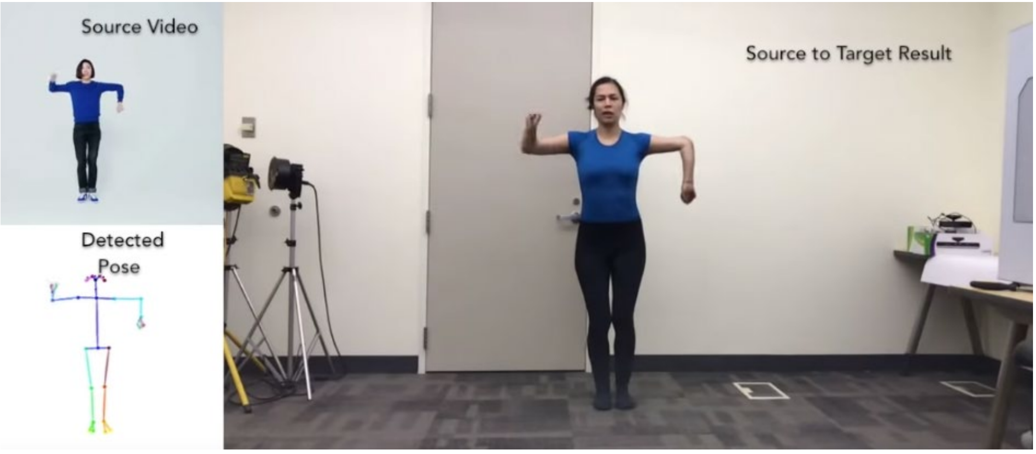
Demonstration of how the identity of a person in a source picture (top row) is transferred to a target (middle row), resulting in a face swap (bottom row).[50]

Figure 5 - Face generation



Demonstration by StarGAN v2 by NAVER Corporation (Creative Commons BY-NC 4.0 license) in which an input picture (left) is transformed into celebrity-like photographs.[51]

Figure 6 - Full body puppetry from a demo video called 'Everybody Dance Now' (2018)



The algorithm detects the pose of a professional dancer in a source video (top left) and generates a video (right) in which a target person makes the same dance moves.[52]

[50] Li et al., 'FaceShifter.'

[51] Choi et al., 'StarGAN.'

[52] Caroline Chan, *Everybody Dance Now*, 2018.

3.3. Voice cloning technology

Voice cloning technology enables computers to create an imitation of a human voice. Voice cloning technologies are also known as audio-graphic deepfakes, speech synthesis or voice conversion/swapping.[53] AI voice cloning software methods can generate synthetic speech that is remarkably similar to a targeted human voice. Some believe that the difference between a real and a synthesised voice is becoming 'imperceptible to the average person'.[54]

The development of AI voice cloning software began decades ago when a number of methods were invented for computers to synthesise voice. These so-called Text-to-Speech (TTS) algorithms are able to convert text into spoken words. This allowed computers to use voice for interacting with humans. In many cases - such as announcement systems in train stations - traditional audio messages have been replaced by a TTS system, eliminating the need to pre-record every possible message and offering much greater flexibility.

Traditionally there are two approaches to TTS: Concatenative TTS and Parametric TTS.[55] Concatenative TTS utilises a database of audio clips containing words and sounds that can be combined to form full sentences. The resulting audio is understandable, but has a typical robotic ring to it. It is difficult to express emotions or use subtle intonations in Concatenative TTS which is normal in natural speech. Using Concatenative TTS to clone a voice requires a serious investment, as for every new voice a new database has to be built.

Parametric TTS takes a different approach. Instead of using pre-recorded audio clips it uses a model of a voice. This model can be derived from recordings of a target, and is increasingly able to capture the characteristic sound and subtleties of a person's pronunciation. Once a Parametric TTS system has been built to create a model of a specific target, it can be reused to create models of other targets as well. This greatly reduces the operational costs compared to Concatenative TTS. However, before the invention of modern AI techniques such as GANs (See Annex 3) this method yielded unconvincing results, and humans were able to quickly recognise that the resulting audio was an imitation.

Today, Artificial Intelligence (AI) has enormously increased the quality of Parametric TTS-based voice cloning. TTS has become a standard feature of everyday consumer electronics. Popular TTS-based devices are voice assistants, such as Google Home, Apple Siri and Amazon Alexa and navigation systems.

The barriers to creating voice clones are reducing due to a variety of easily accessible and reusable AI-powered tools such as Tacotron[56], WaveNet[57], Deep Voice[58], or Voice Loop.[59] These systems are capable of imitating the sound of any person's voice, and can 'pronounce' a text input. An audio clip with just a few minutes of recorded speech can already be enough to extract the characteristic features of a person's voice. The extracted information is used to create an AI voice model. Based on this model a computer can generate new audio clips in which any text could be pronounced with a sound that is

[53] Medikonda Neelima and I Santiprabha, 'Mimicry Voice Detection Using Convolutional Neural Networks,' in *2020 International Conference on Smart Electronics and Communication (ICOSEC)* 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India: IEEE, 2020, 314–18.

[54] Kim Martin and V. P. Marketing, 'What Is AI Voice Cloning Software? Find Out at ID R&D,' *ID R&D* (blog), March 9, 2020.

[55] Sciforce, 'Text-to-Speech Synthesis: An Overview,' Medium, February 13, 2020.

[56] Jonathan Shen et al., 'Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,' *ArXiv:1712.05884 [Cs]*, February 15, 2018.

[57] Aäron van den Oord and Sander Dieleman, 'WaveNet: A Generative Model for Raw Audio,' Deepmind, accessed January 25, 2021.

[58] Sercan O. Arik et al., 'Deep Voice: Real-Time Neural Text-to-Speech,' *ArXiv:1702.07825 [Cs]*, March 7, 2017.

[59] Yaniv Taigman et al., 'VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop,' *ArXiv:1707.06588 [Cs]*, February 1, 2018.

very similar to the target voice. For example, as part of an advertisement campaign for snacks, users can generate a custom video in which Argentine football celebrity Lionel Messi seems to speak English fluently.[60]

The quality of the output by AI-based TTS systems are steadily improving. Nowadays, the models are able to learn based on the discovery of new patterns in audio data. The invention of GANs - which are also pivotal to the acceleration of graphic deepfakes (See Annex 3 for a detailed description of GANs) - has also accelerated the development of voice clones, resulting in increasingly convincing clones that are harder to detect by humans.

Thus, the use of AI technology gives a new dimension to clone credibility and the speed at which a credible clone can be created. However, it is not just the sound of a voice that makes it a convincing clone. The content of the audio clip also has to match the style and vocabulary of the target. Voice cloning technology is therefore connected to the next paragraph on text synthesis technology, which can be used to automatically generate content creation that resembles the target's style.

3.4. Text synthesis technology

Text synthesis technology is used in the context of deepfakes to generate texts that imitate the unique writing and speaking style of a target. The technologies lean heavily on Natural Language Processing (NLP); a scientific discipline at the intersection of computer science and linguistics. Its primary application is to improve textual and verbal interactions between humans and computers.

NLP systems can analyse large amounts of texts, including transcripts of audio clips of a particular target. This results in a system which is capable of interpreting a speech to some extent, including the words as well as a level of understanding of the emotional subtleties and intentions expressed.. This can result in a model of a person's speaking style, which can in turn be used to synthesise novel speeches.

Common architecture used in NLP is a deep learning algorithm called the Transformer. This algorithm is basically able to 'transform' an input text into a new text, by learning how the sequence of words relate to each other in sentences and texts. One of the most advanced in a series of language models built on this architecture is Generative Pre-trained Transformer 3 (GPT-3)[61] created by OpenAI, a San Francisco-based artificial intelligence research laboratory. GPT-3 is a general-purpose NLP that has showed impressive performance with translation, question-answering, as well as with unscrambling words. The OpenAI researchers claim that 'GPT-3 can even generate news articles which human evaluators have difficulty distinguishing from articles written by humans'. At present, large amounts of computing power, electricity and training data are needed to create GPT-3 models. This has led to scrutiny by prominent AI ethics researchers on the environmental impact of this technology.[62] However, the OpenAI researchers state that once such a model is trained, it takes relatively low-power computers to use the model and generate large amounts (hundreds of pages) of text.

3.5. Trends in deepfake videos, voice cloning and text synthesis

Since the inception of the term deepfakes less than five years ago, the concept itself and its predecessors have developed rapidly. There are a number of key drivers that have enabled a number of trends.

[60] The service can be found on https://www.messimessages.com

[61] Ram Sagar, 'OpenAI Releases GPT-3, The Largest Model So Far,' *Analytics India Magazine* (blog), June 3, 2020; Robert Dale, 'GPT-3: What's It Good for?,' *Natural Language Engineering* 27, no. 1 (2021): 3.

[62] Karen Hao, 'We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says.,' MIT Technology Review, 2020.

The key drivers are:

- **Availability of datasets and computing power**. The computer vision community has created large datasets with labelled visual material, and many of these are freely available on the internet. These datasets are necessary for training the machine learning algorithms. The creators of deepfakes can readily access these datasets, eliminating the time-consuming work of gathering and labelling material. Moreover, the required computing power for training machine learning algorithms is available at low cost due to cloud computing services.[63] Some services, such as Google Colab, actually provide enough computing power for creating short high-quality deepfake videos in a matter of hours. When utilising multiple Google accounts, it is possible to gain access to a significant amount of computing power at zero monetary costs. Thus, a regular computer or even a smart phone with internet access suffices for creating high-quality deepfakes.
- **Accessibility of high-quality algorithms and pre-trained models**. The academic community is accustomed to publishing work in open or easily accessible journals and code repositories, such as Github. This drives a strong uptake tendency by the creators of deepfake software. Additionally, pre-trained machine learning models are shared among deepfake creators. Models only need to be trained once and can be reused indefinitely, eliminating a time-consuming step of training models on datasets and eliminating partially the need for computing power.
- **5G connectivity.** Across Europe telecom operators are launching the next generation of mobile connectivity networks. These 5G networks offer increased bandwidth, enabling users to stream and view video content at higher qualities as well as use portable virtual and augmented reality systems.
- **Rise of 3D sensors.** The latest generation of consumer electronics are equipped with 3D sensors. At first, these were mainly used for authentication purposes, such as unlocking smart phones by scanning the user's face. The latest Apple iPhone and iPad now also contain general purpose 3D sensors that can be used to capture 3D information of entire scenes and scan objects. It is expected that the creators of deepfakes will soon benefit from obtaining large quantities of 3D data on their targets' faces.
- **Cat-and-mouse game between producers and detectors.** Paradoxically, increased image forensics and deepfake detection capabilities drive towards increased quality of deepfake videos. As described in the section on GANs (Annex 3), the algorithms that create deepfakes benefit from detectors due to the learning capacity based on feedback loops.[64] This also explains why many of the scholarly articles on deepfake detection are published by the same authors that work on algorithms that create deepfake capabilities. This innovation cycle is further catalysed by the availability of shared libraries of deepfake videos, which are supplemented frequently with the products of the latest deepfake creation algorithms, and used to develop and benchmark new detection methods.[65]

These drivers lead to a number of trends:

- **Live real-time deepfakes.** The additional bandwidth offered by new communication technologies such as 5G enable 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.

[63] Henry Ajder, 'The State of Deepfakes: Landscape, Threats, and Impact' Sensity, 2019.

[64] Cade Metz and Keith Collins, 'How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos,' *The New York Times*, January 2, 2018, sec. Technology.

[65] Andreas Rössler et al., 'FaceForensics++: Learning to Detect Manipulated Facial Images,' *ArXiv:1901.08971 [Cs]*, August 26, 2019.

- **Supply and demand platforms for deepfakes.** The strong media appeal and increased popularity of video media have created a market for manipulated videos that are facilitated by supply and demand platforms. There are special marketplaces on which users or potential buyers can post requests for deepfake videos. For example, requests for non-consensual pornography videos of celebrities are fulfilled on internet forums, and certain websites are dedicated to sharing such videos.
- **Commodification of deepfake tools**. The availability of computing power and accessibility of high-quality algorithms lead to a rapid commodification of deepfake tools. Advanced deepfake software suites are freely distributed and accompanied by instructional materials, making it relatively easy for those with some background in computer programming to get started.[66] Software suites for video manipulation also offer marketplaces for exchanging deepfake algorithms.[67] Moreover, several easy-to-use smart phone applications exist, that require no technical know-how whatsoever.[68] There are even chat bots on platforms like Telegram that return a deepfake to anyone that sends them an image; disturbingly and notoriously known for virtually undressing women, including under-age victims.[69]
- **Deepfake as a service companies.** The increased demand for deepfakes has also led to the established of several companies that deliver deepfakes as a product or even online service. On platforms like Synthesia and Rephrase anyone can generate videos, based on text input and a target video. These services are intended for use by marketers to personalise videos, eliminating the need to record a video for each recipient. Essentially, these services make producing a deepfake video as easy as editing text.
- **AI and 3D animation hybrids.** The advent of photorealistic 3D avatar technology offers clear synergetic opportunities when combined with AI-based deepfake technology. There are already publications and services on the market that show that deepfake creators combine both approaches (See Figure 7).[70]
- **Reduced input requirements.** There is a trend among deepfake creators to develop algorithms that can generate high-quality output, based on very little input. For example, some algorithms seem capable of generating deepfake videos based on a single picture of the target, or generate audio speeches that convincingly resemble the target's voice based on only a few seconds of audio.[71] This means that the availability of large quantities of visual data of a particular person is no longer a requirement, making anyone with only a small number of audio-visual representations on the internet a potential target.

[66] dfaker, 'Df'; iperov, 'DeepFaceLab'; Deepfake, 'Faceswap.'

[67] 'Runway | Make the Impossible,' Runway, accessed May 4, 2021.

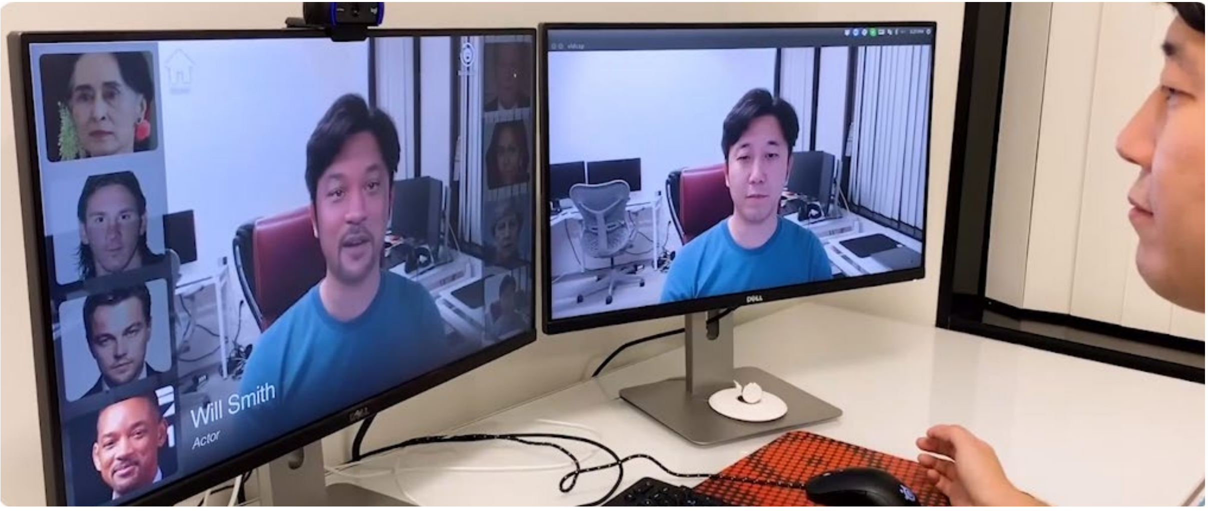
[68] For example: FakeApp, FaceApp, Zao, DeepNude, etc

[69] Giorgio Patrini, 'Automating Image Abuse: Deepfake Bots on Telegram,' *Sensity* (blog), October 20, 2020.

[70] Koki Nagano et al., 'PaGAN: Real-Time Avatars Using Dynamic Textures,' *ACM Transactions on Graphics* 37, no. 6 (December 4, 2018): 258:1-258:12; Jie Cao et al., '3D Aided Duet GANs for Multi-View Face Image Synthesis,' *IEEE Transactions on Information Forensics and Security* 14, no. 8 (August 2019): 2028–42; Zejian Wang et al., 'AI-Synthesized Avatars: From Real-Time Deepfakes to Virtual AI Companions,' in *ACM SIGGRAPH 2020 Real-Time Live!*, SIGGRAPH '20 Virtual Event, USA: Association for Computing Machinery, 2020, 1.

[71] Egor Zakharov et al., 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models,' *ArXiv:1905.08233 [Cs]*, September 25, 2019.

Figure 7 - Demonstration of real-time deepfake technology combining 3D animation and AI techniques



The screen on the right shows the webcam input, the screen on the left is the manipulated output also displaying other pre-trained models, from which the user can choose in this software by Pinscreen Inc (2020)[72]

3.5.1. Five-year future scenario and risk development

When projecting the trends and drivers described above into the future, a scenario starts to form. Most likely the tools for creating deepfakes will become abundantly available and easy to use within a matter of years. Already we see that smartphone apps that unlock only a part of the potential of the technology quickly become wildly popular. FaceApp for example allows users to change their images, such as appearing older. It was downloaded over 150 million times in mid 2019.[73] In 2021 the app Wombo that applies lip-sync technology to images in order to create satiric videos was downloaded over 2 million times in the first two weeks after its release.[74] Therefore, it is expected that the functionalities of these apps will be adopted by mainstream software and become part of the everyday use of social media within the next five years. The current rise of deepfake-as-a-service companies, and the uptake by large corporations like SAP[75], means deepfake videos and audio will be commonly used in software products and games. This mainstreaming effect of the technology means that a large part of the European population will become familiar with the technology in the near future.

The expected sharp increase in availability will also translate into a much higher likelihood of abuse. Whereas today there are only few examples of high-profile incidents linked to deepfake techniques, such as the attempted coup-d'etat in Gabon, and non-consensual deepfake pornography mainly targeted at female celebrities, this will likely become more widespread.

Preventative strategies, such as raising awareness of the existence of deepfakes and filtering of nefarious deepfakes by social media platforms, will reduce some of the potential impact. However, given the cat-and-mouse-game dynamic between deepfake creators and detectors, it is likely that advanced actors will still be able to create undetectable forgeries and mislead their targets.

[72] 'Unreal PaGAN: AI-Generated Real-Time Avatars in UE4,' Pinscreen, 2020.

[73] Mansoor Iqbal, 'FaceApp Revenue and Usage Statistics (2020),' Business of Apps, September 5, 2019.

[74] Steven Asarch, 'Wombo.Ai Lets Users Make Silly Deepfake Videos of Their Friends or Celebrities Singing Songs,' Insider, 2021.

[75] 'How to Use AI Generated Photos | Generated.Photos,' accessed May 4, 2021.

Thus, within the next five years, the nefarious use of deepfake technology will probably develop from a high impact, but low likelihood risk, into a high impact with moderate to high likelihood risk.

At the same time, the lowering of barriers for the use of deepfake technology will also catalyse its use for beneficial purposes. For example, it is likely people will more often encounter life-like avatars that serve as virtual assistants. The current virtual assistants such as Google Home and Amazon's Alexa might be extended with a screen on which a human-like image is visible, which creates the illusion of having a conversation with a (familiar) person instead of a robot.

The integration of deepfake technology in augmented reality systems may introduce new risks, that are not yet understood. Suppose a user has selected an avatar with the voice of a relative for the presentation of news from sources the user has selected. This could lead to a scenario in which a trusted person seems to pronounce disinformation. Although the psychological effects of having a trusted person present disinformation are not yet understood, it could be expected that this opens up new avenues for manipulation and accompanying risks.

3.6. Detection software and technical prevention strategies

Public concern about the potential risks of deepfakes has created a demand for detection and prevention.[76] Detection systems are necessary whenever manipulated materials are used as evidence, for example in court and insurance cases[77] or news reporting.[78] Prevention is also necessary, as it has been proven difficult to correct false information once the public has been exposed to it.[79]

In the following section, we will discuss common detection approaches, their limitations, and several common technical prevention strategies.

3.6.1. Detection technology

There are two distinct approaches to deepfake detection: manual and automatic detection.[80] Manual detection requires a skilled person to inspect the video material and look for inconsistencies or cues that might indicate forgery. Another logical approach that some have attempted to automate is to compare other audio-graphic material of the same event.[81] A manual approach could be feasible when dealing with low quantities of suspected materials. However, this approach is not compatible with the scale at which audio-visual materials are used in modern society. Therefore, it is not a feasible solution at a societal level.

Automatic detection software can be based on a (combination of) detectable giveaways:

- **Speaker recognition.** Recognition is based on both identification and verification. A speaker identification system can be used to determine who the speaker is, with just audio as an input. An automatic speaker verification (ASV) system verifies if the voice of a speaker matches the

[76] Chesney and Citron, 'Deep Fakes.'

[77] 'Truepic | Photo and Video Verification Platform,' accessed November 26, 2020.

[78] Nguyen et al., 'Deep Learning for Deepfakes Creation and Detection'; Anushree Deshmukh and Sunil B. Wankhade, 'Deepfake Detection Approaches Using Deep Learning: A Systematic Review,' in *Intelligent Computing and Networking*, ed. Valentina Emilia Balas et al., Lecture Notes in Networks and Systems Singapore: Springer, 2021, 293–302; Shruti Agarwal et al., 'Protecting World Leaders Against Deep Fakes,' 2019, 38–45; Siwei Lyu, 'DeepFake Detection: Current Challenges and Next Steps,' *ArXiv:2003.09234 [Cs]*, March 11, 2020.

[79] Jonas De Keersmaecker and Arne Roets, ''Fake News': Incorrect, but Hard to Correct. The Role of Cognitive Ability on the Impact of False Information on Social Impressions,' *Intelligence* 65 (November 1, 2017): 107–10.

[80] Dafeng Gong, 'Deepfake Forensics, an AI-Synthesized Detection with Deep Convolutional Generative Adversarial Networks,' *International Journal of Advanced Trends in Computer Science and Engineering* 9, no. 3 (June 25, 2020): 2861–70.

[81] Eleanor Tursman et al., 'Towards Untrusted Social Video Verification to Combat Deepfakes via Face Geometry Consistency,' in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, 2784–93.

claimed identity. These technologies are often based on comparing new audio fragments to previously determined voice prints in a database.[82]

- **Voice liveness detection.** This technology is able to detect whether the sound of a voice comes from a live person that is speaking, or a pre-recorded clip. Even when voice clones are indistinguishable to the human ear, these kind of (AI-based) tools can detect artefacts that are not present in the sound of a live voice.[83] These technologies are still applicable, yet become less reliable when the quality of the audio is reduced, such as low quality telephone conversations or radio interviews.
- **Facial recognition**. Software that is used to identify people in photographic materials can also be applied to suspected forged materials.[84] Deepfake algorithms often stretch or wrap faces[85], or only adjust distinct features when creating morphs[86], resulting in irregularities. Whenever facial recognition software fails to identify the person that is claimed to be portrayed it might be an indication of forgery.
- **Facial feature analysis.** Researchers are developing algorithms for practically all facial landmarks, such as the position and movement of the nose, mouth and eyes, to spot artefacts caused by deepfake manipulations. The scientific literature on image forensics, for example, contains papers describing deepfake detectors that analyse the lack of eye-blinking[87] or recognise manipulated eyebrows.[88]
- **Temporal inconsistencies.** Since deepfake videos are often created by modifying each frame of a movie separately, detectable inconsistencies may occur between frames.[89] For example, deepfake algorithms could cause sudden changes in head pose[90], inconsistent lip movement[91] or other unnatural movements of facial landmarks.
- **Visual artefacts.** Whenever an image is partially modified the deepfake algorithm must somehow create a transition between the original material and the manipulation. This often results in a blurry area[92], for example between the object in the foreground and its background.[93]

[82] Ravika Naika, 'An Overview of Automatic Speaker Verification System,' in *Intelligent Computing and Information and Communication*, ed. Subhash Bhalla et al., Advances in Intelligent Systems and Computing Singapore: Springer, 2018, 603–10.

[83] Shang Jiacheng, Si Chen, and Jie Wu, 'Defending Against Voice Spoofing: A Robust Software-Based Liveness Detection System,' in *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Chengdu: IEEE, 2018.

[84] Jian Wu et al., 'A Forensic Method for DeepFake Image Based on Face Recognition,' in *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, HPCCT &amp; BDAI 2020 Qingdao, China: Association for Computing Machinery, 2020, 104–8.

[85] Yuezun Li and Siwei Lyu, 'Exposing DeepFake Videos By Detecting Face Warping Artifacts,' *ArXiv:1811.00656 [Cs]*, May 22, 2019.

[86] Lingzhi Li et al., 'Face X-Ray for More General Face Forgery Detection,' *ArXiv:1912.13458 [Cs]*, April 18, 2020.

[87] Yuezun Li, Ming-Ching Chang, and Siwei Lyu, 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking,' *ArXiv:1806.02877 [Cs]*, June 11, 2018; Tackhyun Jung, Sangwon Kim, and Keecheon Kim, 'DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern,' *IEEE Access* 8 (2020): 83144–54.

[88] Hoang Mark Nguyen and Reza Derakhshani, 'Eyebrow Recognition for Identifying Deepfake Videos,' in *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020, 1–5.

[89] Irene Amerini and Roberto Caldelli, 'Exploiting Prediction Error Inconsistencies through LSTM-Based Classifiers to Detect Deepfake Videos,' in *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, IH&amp;MMSec '20 Denver, CO, USA: Association for Computing Machinery, 2020, 97–102.

[90] Xin Yang, Yuezun Li, and Siwei Lyu, 'Exposing Deep Fakes Using Inconsistent Head Poses,' *ArXiv:1811.00661 [Cs]*, November 13, 2018.

[91] Mousa Tayseer Jafar et al., 'Forensics and Analysis of Deepfake Videos,' in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, 053–058.

[92] Mohammed Akram Younus and Taha Mohammed Hasan, 'Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform,' in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 2020, 186–90.

[93] Weiguo Zhang, Chenggang Zhao, and Yuxing Li, 'A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis,' *Entropy* 22, no. 2 (February 2020): 249.

Also, when there are only few source materials, the algorithm might have to guess what certain expressions look like on a target's face. In all of these and similar cases, the algorithm may leave detectable artefacts in the output. These artefacts or patterns of artefacts can be detected by algorithms.[94] Also, deepfake algorithms that are trained to generate synthetic images often fail to deliver realistic backgrounds.[95] And accounting for changes in illumination still proves to be a challenge to some deepfake algorithms.

- **Lack of authentic indicators**. Camera sensors consist of tiny pixels, that vary slightly in sensitivity due to the manufacturing process of the chips. These variations result in a sort of noise watermark that can be detected in every image or video that is made with a camera. Deepfake algorithms often disrupt this detectable pattern, which is an indication of forgery.[96]

3.6.2. Detection limits

The extensive literature on deepfake detection methods might look reassuring, but there are several important cautions that need to be kept in mind.[97]

First, the performance of detection algorithms is often measured by benchmarking against a common data set with known deepfake videos, such as the FaceForensics++ database which contains 1.8 million samples.[98] However, a high confidence level in discriminating videos from such a dataset with known deepfakes does not guarantee a trustworthy performance on entirely new materials. In practice, it turns out that detectors are often good at spotting one kind of deepfake.[99] Studies into detection evasion show that even simple modifications can drastically reduce the reliability of a detector.[100]

Another problem detectors face is that audio-graphic material is often compressed or reduced in size when shared on online platforms such as social media and chat apps. The reduction in the number of pixels and artefacts that sound and image compression create can interfere with the ability to detect deepfakes.[101] Also, smartphone camera apps often have enabled filters by default, automatically modifying every image or video, nullifying the very notion of the existence of an authentic image in the first place.

[94] Ali Khodabakhsh and Christoph Busch, 'A Generalizable Deepfake Detector Based on Neural Conditional Distribution Modelling,' in *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2020, 1–5; Akash Chintha et al., 'Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection,' *IEEE Journal of Selected Topics in Signal Processing* 14, no. 5 (August 2020): 1024–37; Mengnan Du et al., 'Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder,' *ArXiv:1909.05999 [Cs]*, September 19, 2020.

[95] Mohammed A. Younus and Taha M. Hasan, 'Abbreviated View of Deepfake Videos Detection Techniques,' in *2020 6th International Engineering Conference 'Sustainable Technology and Development' (IEC)*, 2020, 115–20.

[96] Xu Chang et al., 'DeepFake Face Image Detection Based on Improved VGG Convolutional Neural Network,' in *2020 39th Chinese Control Conference (CCC)*, 2020, 7252–56; Mo Chen et al., 'Determining Image Origin and Integrity Using Sensor Noise,' *IEEE Transactions on Information Forensics and Security* 3, no. 1 (March 2008): 74–90.

[97] Sakshi Agarwal and Lav R. Varshney, 'Limits of Deepfake Detection: A Robust Estimation Viewpoint,' *ArXiv:1905.03493 [Cs, Math, Stat]*, May 9, 2019.

[98] Rössler et al., 'FaceForensics++.'

[99] Zhaohe Zhang and Qingzhong Liu, 'Detect Video Forgery by Performing Transfer Learning on Deep Neural Network,' in *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, ed. Yong Liu et al., Advances in Intelligent Systems and Computing Cham: Springer International Publishing, 2020, 415–22.

[100] Yihao Huang et al., 'FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction,' *ArXiv:2006.075 3 3 [Cs]*, August 17, 2020; Nicholas Carlini and Hany Farid, 'Evading Deepfake-Image Detectors with White- and Black-Box Attacks,' *ArXiv:2004.00622 [Cs]*, April 1, 2020; Shehzeen Hussain et al., 'Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples,' *ArXiv:2002.12749 [Cs]*, November 7, 2020.

[101] Zhang Hongmeng et al., 'A Detection Method for DeepFake Hard Compressed Videos Based on Super-Resolution Reconstruction Using CNN,' in *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, HPCCT &amp; BDAI 2020 Qingdao, China: Association for Computing Machinery, 2020, 98–103.

Automatic speech verification (ASV) systems, which are an active field of research and development, also have serious shortcomings. ASV systems are good at dealing with classic forms of attacks, such as replay and impersonation by another human actor. However, ASV systems are less effective against AI-based attacks[102] and need to increase the use of AI in order to improve detection capabilities.[103]

The most pressing need is to develop a uniform forgery assessment methodology. Current detection systems which are based on only one countermeasure will not suffice. Voice cloning technology will continue to progress. Therefore monitoring technological progress and continuous integrations into a holistic and efficient detection system are needed.[104]

3.6.3. Technical prevention strategies

There are several technical strategies that may prevent an image or audio clip from being used as an input for creating deepfakes or may limit its potential impact. Prevention strategies include adversarial attacks on deepfake algorithms, and strengthening the markers of authenticity of audio-visual materials and technical aids for people to more easily spot deepfakes. In this section, each strategy is described in more detail.

Adversarial attacks on deepfake algorithms are methods that exploit vulnerabilities in computer vision algorithms. It is more or less the digital equivalent of a person wearing makeup to prevent identification by facial recognition cameras. The technology works by adding specific noise patterns to images as an overlay. The overlays are indistinguishable to the human eye. Computer vision algorithms, however, detect the noise and can be fooled into believing these are real features of the image, which will for example hamper the ability to correctly detect an object. This approach has been demonstrated as effective against some deepfake algorithms.[105] However, applying these techniques often requires the attacker to have some knowledge about the detector algorithm, in order to create effective deceptive overlays. Therefore, it is not suitable as a generic prevention against manipulation.

The strengthening of authenticity markers of audio-visual content is often based on somehow registering authentic content or (digitally) watermarking audio-visual materials. Some argue that blockchain or distributed ledger technology (DLT) could be used to register original materials or a unique identifier.[106] Just as the British company Provenance aims to increase the transparency of product supply-chains by registering the origin of every ingredient or component in a blockchain database, a similar system could be created for the supply of audio-graphic information. However, these initiatives often overlook that this solution introduces many new vulnerabilities, such as attacks on the integrity of the DLT itself, or the dependency on technicians and organisations that will be responsible for operating such a system. Also, for these solutions to be effective, there must be a link between the register and the recipient of information, which is not very feasible given the enormous number of devices and software people use to consume audio-graphic materials. Despite these difficulties, some initiatives are exploring the implementation of this approach.[107]

[102] Ian Goodfellow et al., 'Attacking Machine Learning with Adversarial Examples,' OpenAI, February 24, 2017.

[103] Hafiz Malik and Raghavendar Changalvala, 'Fighting AI with AI: Fake Speech Detection Using Deep Learning' Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics, Audio Engineering Society, 2019.

[104] Madhu R. Kamble et al., 'Advances in Anti-Spoofing: From the Perspective of ASVspoof Challenges,' *APSIPA Transactions on Signal and Information Processing* 9 (ed 2020).

[105] Chin-Yuan Yeh et al., 'Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks,' in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2020, 53–62.

[106] Haya R. Hasan and Khaled Salah, 'Combating Deepfake Videos Using Blockchain and Smart Contracts,' *IEEE Access* 7 (2019): 41596–606.

[107] 'Content Authenticity Initiative,' Content Authenticity Initiative, accessed May 6, 2021.

Another approach that could authenticate audio-visual materials is embedding a (digital) watermark in the audio or graphic file itself. 108 This could even be implemented in camera chips. Camera chips already have a unique variation in pixel sensitivity that results in a detectable noise pattern.

Finally, harm could be prevented by supporting people to spot deepfakes. Some initiatives aim to raise awareness and build such capacity by conducting 'prebunking' interventions. This approach entails exposing people to clearly labelled potential deepfakes . Often, malicious deepfake creators follow a common pattern, such as continuously repeating a certain narrative or targeting a particular individual. By informing people about the existence of such misinformation, they may become more critical and resilient when confronted with such videos. This approach has been shown to reduce susceptibility to traditional (non-audio-graphic) misinformation.[109] Several institutions have also developed training software purposely built with this aim.[110]

[108] Adnan Alattar, Ravi Sharma, and John Scriven, 'A System for Mitigating the Problem of Deepfake News Videos Using Watermarking,' *Electronic Imaging* 2020, no. 4 (January 26, 2020): 117-1-117–10.

[109] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren, 'Prebunking Interventions Based on 'Inoculation' Theory Can Reduce Susceptibility to Misinformation across Cultures,' *Harvard Kennedy School Misinformation Review* 1, no. 2 (February 3, 2020).

[110] Matt Groh, 'Project Overview ‹ Detect DeepFakes: How to Counteract Misinformation Created by AI,' MIT Media Lab, 2020.

4. Societal context

4.1. Relevant factors and trends

Media manipulation and doctored imagery are by no means new phenomena. In that sense, deepfakes can be seen as just a new technological expression. But that perspective would fall short when it comes to understanding its potential societal impact. To fully grasp the significance of deepfakes, we need to analyse the societal factors and trends that shape its societal value and the way it becomes ingrained in social life. In this chapter, we will cover relevant factors and trends that play a role in shaping the societal impact of deepfakes.

Changing media landscape

In modern society, media play a pivotal role in the flow of information. According to Kalpokas (2020) media not only '*mediate* between the world and our experience of it, but also *generate* that experience'.[111] Although the idea that media are central to everyday life has become commonplace, he states that the changing nature of media itself is commonly overlooked.[112] Our 'information ecosystem' has changed dramatically as a result of the introduction of the internet.[113] Most notable is a trend that some would describe as the 'democratisation' of media; the growing opportunities for individuals to *be* the media. Although the term 'democratisation' seems to hint at developments in favour of democracy, research shows that user-generated content platforms also involve threats to democracy.[114]

Today, people are no longer solely informed by established news and media organisations, which used to control the quality and speed of information. In the age of the information society, the traditional model of centralised information distribution has been accompanied by direct communications modes via online platforms on a global scale. Consequently, interviewed experts point out that our information ecosystem is now characterised by increased speed, information overload, but also a shortened length and fragmentation of news, as seen in Twitter threads for example.

Growing importance of visual communication

Another important trend is the fact that communication, especially online, is becoming increasingly visual.[115] In recent years, media have increasingly produced visual content.[116] Today, visual communication is the norm rather than the exception.[117] The explanation for the growing popularity of visuals as a dominant mode of communication is the fact that it is a very effective way of transmitting information. Images are effective communication means, because audiences can create and retrieve memories more easily when exposed to visuals.[118] Therefore, audio-visual media have a strong appeal for their audience and may have a unique psychological power.

Growing spread of mis-, dis- and malinformation

At first, the participation of citizens in the creation and distribution of news on social media was seen as a major boost for democratising processes, and as a welcome opportunity to overcome distorted

[111] Kalpokas, 'Problematising Reality.'

[112] Kalpokas.

[113] Schick, *Deep Fakes and the Infocalypse*.

[114] Pieter van Boheemen, Geert Munnichs, and Elma Dujso, 'Digital Threats to Democracy' The Hague: Rathenau Instituut, 2020.

[115] Kalpokas, 'Problematising Reality.'

[116] Erin McCoy, 'Visual Communication Is Transforming Marketing – Are You Up To Speed?,' Forbes, 2017.

[117] Ana Costa, Joost Bakker, and Gabriela Plucinska, 'How and Why It Works: The Principles and History behind Visual Communication,' *Medical Writing* 29 (March 1, 2020): 16–21.

[118] Vaccari and Chadwick, 'Deepfakes and Disinformation.'

representations of the truth by corrupt governments and their own media stations. However, over time it has become clear that social media has also made it more difficult to screen and control content, and harder (if not impossible) to agree on ethical or professional codes of conduct, or norms and behaviours. As social media grew, the spread of several kinds of misleading information grew as well.

The Council of Europe classified misleading information into the classes of mis-, dis- and malinformation:[119]

- Dis-information: Information that is false and deliberately created to harm a person, social group, organisation or country.
- Mis-information: Information that is false, but not created with the intention of causing harm.
- Mal-information: Information that is based on reality, used to inflict harm on a person, organisation or country.

Internationally, disinformation campaigns have become widespread.[120] Given the fact that visualisation is an easy way to transmit information, and thus also disinformation, visual media play an important role in the growing spread of disinformation.

4.2. Welcoming environment for deepfakes

The changing media landscape, together with the dominance of visualisation as a communication strategy, creates a welcoming environment for deepfakes. But there are other explanations for the increasing popularity of deepfakes. Deepfakes are still a novelty to many and often attract media attention due to the sensational nature of the videos, such as parodies[121] and non-consensual pornography. Mainstream media also frequently use deepfake technology in prime-time broadcasts, such as the British Channel 4 parody on the traditional Christmas message of Queen Elizabeth II in December 2020.[122] In this speech – which was broadcast at the same time as the official Christmas message and clearly labelled as deepfake– the Queen seems to refer to controversial topics and performs a TikTok dance challenge. According to Channel 4 it was intended as a warning on 'fake news'. The media appeal of such parodies has driven the demand for deepfake videos and has led to the establishment of firms specialised in delivering satiric deepfake videos.

Simultaneously, we can see an increase in popularity of manipulated video media and mixed reality**.** Video media services and platforms are among the most popular internet services. Visual-first social media platforms such as SnapChat, TikTok and Instagram are on the rise, especially among young people. Live-streaming video is also steadily becoming mainstream, as content on existing platforms like Instagram, Facebook and Reddit, and on dedicated platforms like Twitch. Due to the COVID19 lockdowns, the use of streaming video-conferencing tools has also sharply increased. All video media offer some sort of manipulation tools, such as virtual background, face filters and video editing tools. This drives the normalisation of mixed reality experiences.[123]

[119] Claire Wardle and Media Derakhshan, 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking,' *Shorenstein Center*, October 31, 2017.

[120] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy.'

[121] Trey Parker, Matt Stone, and Peter Serafinowicz, *Sassy Justice with Fred Sassy (Full Episode)*, 2020.

[122] 'Deepfake Queen to Deliver Channel 4 Christmas Message,' *BBC News*, December 23, 2020, sec. Technology.

[123] Leonard Yoon et al., 'A Mixed Reality Telepresence System for Dissimilar Spaces Using Full-Body Avatar,' in *SIGGRAPH Asia 2020 XR*, SA '20 Virtual Event, Republic of Korea: Association for Computing Machinery, 2020, 1–2.

4.3. Deepfakes are a catalyst for greater gender inequality

The majority of deepfake videos that are currently circulating online contain sexual images. Deepfake technology has made it relatively easy to 'undress' someone, or swap their face into an already existing pornographic video. Research company Sensity AI estimates that between 90% and 95% of all deepfakes concern non-consensual pornography.[124] The vast majority of those deepfakes, about 90%, are targeted at women. In 2020, the same company reported on a Telegram chat bot that can be used to create deepnude portraits. All users have to do, is provide the chatbot with a picture of someone, who then gets undressed. The bot exclusively creates nudes of women. By the time of the public report, already over 100,000 women were targeted.[125]

These statistics highlight an important problem; the gendered impact of deepfake abuse and exploitation.[126] According to Sam Gregory of the NGO WITNESS, these kind of deepfake applications are 'gendered by design'.[127] The technology that is used to create such content is exclusively programmed for the female body. While previously such videos were mostly targeted at female celebrities, because of the availability of large quantities of data of these individuals, it has over time become easier to make convincing videos of non-famous people. The problem is exacerbated by the fact that deepfake detection systems are biased towards detecting males.[128]As a result, there are growing concerns about the use of deepfake technology for the creation of revenge porn, sextortion and other forms of sexual abuse.

Deepfake pornography not only poses a threat to the rights of individuals, but also to public debate and the functioning of democratic societies, since it can be used to discredit female journalists and politicians too.[129] Deepfakes provide a powerful tool for exacerbating existing gender inequalities and power relations. Forcing women into a virtual sexual context, reduces them to defenceless objects. As such, deepfake pornography and other non-consensual sexual content can be understood as a new form of sexual violence.

4.4. Deepfakes bring a new dimension to disinformation

There are different possible forms of disinformation based on deepfake technologies. First, deepfakes can take the form of convincing misinformation. Fiction may become indistinguishable from fact to an ordinary citizen. Second, disinformation may be complemented with deepfake materials to increase its misleading potential. Third, deepfakes can be used in combination with political micro-targeting techniques. Such targeted deepfakes can be especially impactful. Micro-targeting is an advertising method that allows producers to send customised deepfakes that strongly resonate with a specific audience.

Looking into recent developments in politics and media, the problem of disinformation reveals a very complex challenge. Deepfakes can be considered in the wider context of digital disinformation, alternative facts and changes in journalism. Here, interviewed experts indicate deepfakes are only the tip of the iceberg, or in other words, minor elements of the multifaceted phenomenon of mis- and disinformation shaping current developments in the field of news and media. These comprise phenomena and developments such as alternative facts, fake news, the manipulation of social media

[124] Georgio Patrini, 'Mapping the Deepfake Landscape,' *Sensity* (blog), October 7, 2019.

[125] Patrini, 'Automating Image Abuse.'

[126] Chesney and Citron, 'Deep Fakes.'

[127] Interview Gregory

[128] Jim Nash, 'Bias in Facial Recognition Is Handicapping Deepfake Detection,' Biometric Update, May 17, 2021.

[129] Interview Gregory

channels by trolls or social bots, or even public distrust of scientific evidence.[130] Research by the NGO Avaaz has also demonstrated that the risk of exposure to misinformation depends on the language users speak. Out of an analysis of a sample of English, Italian, Spanish, French, Portuguese, and Arabic misinformation, it appeared that especially Italian and Spanish speaking users are at risk, because detection and labelling seems biased towards the English language[131]. Concerning the specific role of deepfakes in the context of political misinformation, an empirical study has shown that deepfakes can convince people to believe in events that have never occurred, but interestingly, not at a higher rate than with other means of deception such as texts or audio recordings.[132]

4.5. Truth becomes blurry

The rise of deepfakes is likely to have a greater impact than merely individual damage (see Chapter 5). An important societal outcome could be confusion and an erosion of trust. One of the interviewees pointed out that 'the biggest problem of disinformation is that people become confused'. This statement indicates a more general development in public discourse, namely that boundaries between falsehoods and truth are increasingly being blurred.

Uncovering the 'fake' in fake news does not necessarily convince the recipients of its forgery. On the contrary, real news is increasingly distrusted, and proving authenticity and telling the 'truth' becomes an increasing challenge for journalists. A recent empirical study has demonstrated this dynamic for fake political videos. The authors have shown that if recipients know in advance that a video they are watching might be fake, they started to distrust all videos, regardless of whether they were 'real' or 'fake'. Further, the research showed that the recipients were unable to distinguish between an authentic and a faked video. The authors conclude: 'Our findings suggest that even if deepfakes are not themselves persuasive, rhetoric about deepfakes can nevertheless be weaponised by politicians and campaigns to dismiss and disown real videos'.[133]

These developments demonstrate that the rise of deepfakes result in a need to increase efforts for establishing trust. The recent incident in which numerous European Members of Parliament were misleadingly believed to have had a video meeting with Russian opposition leader Navalny's chief of staff Volkov illustrates the urgency.[134] Even though there probably was no deepfake technology at play[135], the incident shows weaknesses in the established verification procedures. The incident shows that the existence of deepfake technology requires a higher level of distrust, which in itself means that the truth becomes more blurry. It basically means that audio-graphic interaction will more often be distrusted, unless proven.

[130] Patrick Gensing, *Fakten gegen Fake News oder Der Kampf um die Demokratie* Duden, 2019.

[131] 'How Facebook Can Flatten the Curve of the Coronavirus Infodemic' Avaaz, 2020.

[132] S Barari, C Lucas, and K Munger, 'Political Deepfake Videos Misinform the Public, But No More than Other Fake Media,' 2021.

[133] John Ternovski, Joshua Kalla, and Peter Aronow, 'Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments,' *OSF Preprints*, 2021.

[134] Andrew Roth, 'European MPs Targeted by Deepfake Video Calls Imitating Russian Opposition,' *The Guardian*, 2021.

[135] James Vincent, ''Deepfake' That Supposedly Fooled European Politicians Was Just a Look-Alike, Say Pranksters,' 2021.

5. Benefits, risks and impacts of deepfakes

Deepfake technologies can be used for a wide variety of purposes, with both positive and negative impacts. In this chapter, the benefits and risks associated with deepfakes are presented based on literature study and expert interviews. We start by describing the beneficial uses of deepfake technology, followed by the different types of risks associated with deepfakes. The chapter concludes with an analysis and overview of the different levels on which deepfakes can have impacts.

5.1. Benefits

Anyone who has used a modern smart phone for photography has probably experienced some benefits of basic deepfake technologies. Often camera apps are equipped with beauty filters, that automatically modify images. More advanced deepfakes in which entire faces are exchanged or speech is modified can also be lawfully created to provide for example critical comments, satire and parodies or simply to entertain an audience. Other obvious possibilities for beneficial use of deepfakes are in the context of audio-graphic productions, human-machine interactions, video conferencing, satire, personal creative expression, and medical treatment or research. In the paragraphs below, the benefits of these applications are described in more detail.

Audio graphic productions

Deepfake technologies first and foremost offer many benefits to audio, photo and video producers and editors. Movie producers can for example fix misspoken lines through voice dubbing, make script changes without the cost of intensive rerecording of footage, or create dubs of actors speaking different languages.[136] [137] Game studios already use deepfake technology to rapidly create numerous 3D models of game characters.[138] In the foreseeable future, movie stars could even use deepfake technologies and share their personal digital models with producers, so that they can create new footage without having to travel to a video shoot. In addition, the use of deepfakes could enable the creation of more realistic stunt doubles and the adaption of the age of actors, to look older or younger.

Deepfake technologies can be used to automatically create a large number of variations of the same audio-graphic material. This is useful in commercial settings, for example to create personalised video message for each and every customer of a company based on a single recording. Commercial companies like Rephrase and Synthesia already offer this as a service.

In addition to editing and making variations, the capability to paste a person into a scene may also serve many other beneficial goals. Deepfake technology can be used to create memorial videos featuring a deceased person, or historic events can be re-enacted.[139] Examples include the virtual revival of historical figures, such as a video for The Dali Museum featuring the famous artist[140], or bringing Michael Jackson back to life in a concert. On the other hand, future scenarios could also be simulated in videos and tested on an audience.[141] Videos of the past and future could both be used in

[136] Jan Kietzmann et al., 'Deepfakes: Trick or Treat?,' *Business Horizons*, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, 63, no. 2 (March 1, 2020): 135–46.

[137] These techniques are already marketed as services, by companies such as Descript https://www.descript.com/

[138] 'AI-Generated Facial Photos For 3D Human Creation | Headshot Plugin | Character Creator,' AI-Generated Facial Photos For 3D Human Creation | Headshot Plugin | Character Creator, accessed May 11, 2021.

[139] Catherine Kerner and Mathias Risse, 'Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds,' *Moral Philosophy and Politics*, November 11, 2020.

[140] 'Dalí Lives (via Artificial Intelligence),' Salvador Dalí Museum, accessed February 17, 2021.

[141] Nicholas Caporusso, 'Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology,' in *Advances in Artificial Intelligence, Software and Systems Engineering*, ed. Tareq Ahram, vol. 1213 Cham: Springer International Publishing, 2021, 235–41.

educational settings.[142] Others see great potential for applications in the advertising and fashion industries, enabling customers to virtually fit clothes by replacing models with their own appearance.[143]

Human-machine interactions

Deepfake technologies can be used to improve digital experiences and interactions between humans and computers.[144] Users may find interacting with an imitation of an acquaintance easier than interacting with a virtual stranger. The company Pinscreen has for example already demonstrated how deepfake technology can be used to improve 3D virtual assistants, enabling more natural human-like interactions.[145] One of the services offered by the deepfakes-as-a-service company Generated Photos allows its customers to use synthetic portraits in their products, without having to worry about copyrights and royalties.[146]

Furthermore, the accessibility of video information can also be greatly improved, by using deepfake technology to automatically translate the content. The company Papercup launched such an automatic translation service. The NGO Malaria No More attracted significant attention when it made a video public in which David Beckham seems to speak nine different languages.[147]

Video conferencing

Video conferencing may also benefit from implementing deepfake techniques. Advanced applications include the possibility for users to be present as life-like avatars in virtual events, such as VR conferences or games.[148] There are also less futuristic applications, such as improving the quality of normal video conferencing. Deepfake technologies could for example greatly reduce the required bandwidth. In a normal video conference, the participants' computers exchange a continuous stream of pictures. Instead, using deepfake technology, they would only need to exchange instructions for each recipient to reconstruct and animate the picture. Such AI-based video compression can also be used to improve the image quality of low-bandwidth video conferences and align faces in such a way that it appears that a person is looking straight at the camera. This could mask what a person is actually looking at, improving the privacy of video conference participants (See Figure 8).[149]

[142] Thies et al., 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos.'

[143] Julia Dietmar, 'GANs And Deepfakes Could Revolutionize The Fashion Industry,' Forbes, 2019.

[144] Andrei O. J. Kwok and Sharon G. M. Koh, 'Deepfake: A Social Construction of Technology Perspective,' *Current Issues in Tourism* 0, no. 0 (March 14, 2020): 1–5.

[145] 'Pinscreen Virtual Assistant (Live Demo 2020) - YouTube,' accessed February 15, 2021.

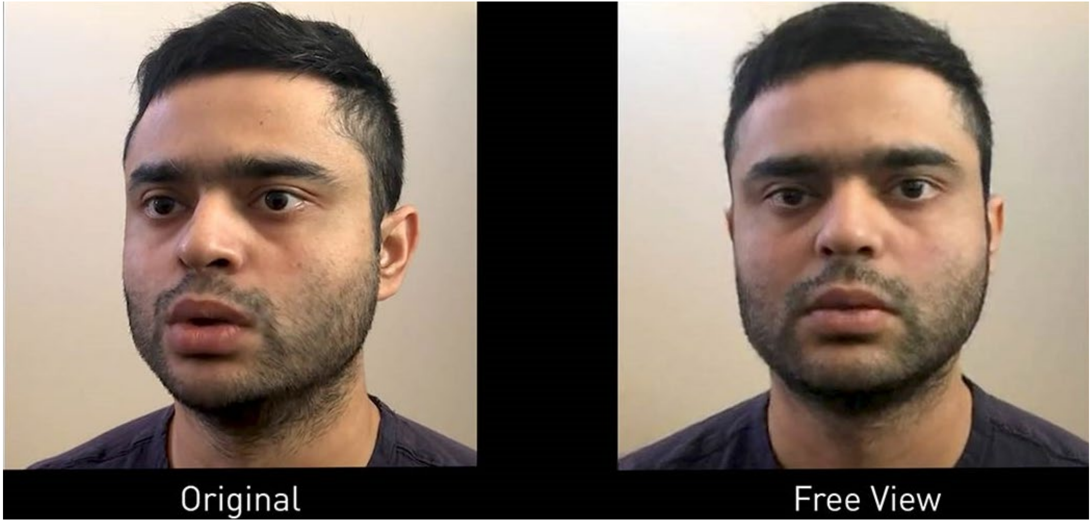
[146] 'How to Use AI Generated Photos | Generated.Photos,' accessed May 11, 2021.

[147] 'Synthesia Insights: Case Study - David Beckham / Malaria No More / RGA,' 2020.

[148] 'Facebook Is Building the Future of Connection with Lifelike Avatars,' Facebook Technology, March 13, 2019.

[149] 'NVIDIA Maxine Video Conferencing Platform,' NVIDIA Developer, October 1, 2020.

Figure 8 - A demonstration of NVidia's face alignment system for video conference using Deepfake technology



Original on the left, aligned image on the right.[150]

Satire

In the interviews, the most often mentioned large scale benign uses of deepfake technologies are parody and satire. Based on a scan of videos listed as deepfakes on video sharing platforms we conclude that many of the existing non-pornographic deepfakes clearly have such a critical or humorous intent. Deepfake satire has already been adopted by mainstream media, such as the Channel 4 Queen Elizabeth II Christmas 2020 speech parody. There are even satiric shows that entirely revolve around footage created by deepfake technology.[151] Not all applications of deepfake technology in parody and satire are benign though: it has been shown that extremist groups can use satirical expressions as a cover for spreading their ideology.[152]

Personal or artistic creative expression

The accessibility and availability of Deepfake technologies also means that people now have new tools to express their creativity. People could feature themselves in Hollywood movies[153], visualise dreams or imaginations about themselves and personalise virtually any picture or video.[154] This variety of self-expressive goals could be seen as a gain in personal or artistic autonomy.[155]

Medical (research) applications

There are beneficial applications of deepfake technology in the field of medical research and treatments. In dentistry and cosmetic surgery for example for the reconstruction of the face after the treatment of cleft palate. Another positive application in medicine is voice creation for hard-of-hearing, deaf or mute people.[156] Face-swapping technology also makes it possible to anonymise portrait videos, hiding research participants' identity in videos while preserving information about their facial

[150] 'NVIDIA Maxine Video Conferencing Platform.'

[151] For example: Parker, Stone, and Serafinowicz, *Sassy Justice with Fred Sassy (Full Episode)*.

[152] Viveca S. Greene, ''Deplorable' Satire: Alt-Right Memes, White Genocide Tweets, and Redpilling Normies,' *Studies in American Humor* 5, no. 1 (2019): 31–69.

[153] Kietzmann et al., 'Deepfakes.'

[154] Chesney and Citron, 'Deep Fakes.'

[155] Chesney and Citron.

[156] 'Snapshot Paper - Deepfakes and Audiovisual Disinformation.'

expressions.[157] There are also experimental therapies to treat anxiety disorders, by showing patients deepfake videos of themselves in which they seem to overcome their fears.[158]

5.2. Risks, harms and impact

Deepfake technologies are enabled by AI technologies and may also have a malicious, misleading and even destructive potential at an individual, organisational, and societal level. The interviewed experts for this research felt that the term 'deepfake' itself has an inherent negative connotation, pointing towards widely-perceived negative impacts and malicious outcomes of deepfake technologies. They say that deepfakes may irritate, humiliate, and even spur violence. The interviewee, Justus Thies, stated that, 'Deepfakes are a poor outgrowth of synthetic media' with a 'malicious strand to exploit AI technology'.[159] Misuse and abuse of deepfake technologies is therefore giving rise to calls for criminalisation. One interviewee explicitly stated that we should not only speak about risks but rather about the dual-use of deepfake technologies. In the strict sense, dual use means a technology can serve civilian and military purposes. In the broader sense, it means it has beneficial and malicious uses. Both meanings apply to deepfake technology.

The broad range of possible risks can be differentiated into three categories of harm: psychological, financial and societal risks.[160] Since deepfakes target individual persons, there are firstly direct psychological consequences for the target. Second, it is also clear that deepfakes are created and distributed with the intent to cause a wide range of financial harms. And thirdly, there are grave concerns about the overarching societal consequences of the technology. An overview of the risks identified in this research are presented and categorised in Table 2 below, and will be described in the following sections.

Table 2 - Overview of different types of risks associated with deepfakes [161]

Psychological harm	Financial harm	Societal harm
• (S)extortion • Defamation • Intimidation • Bullying • Undermining trust	• Extortion • Identity theft • Fraud (e.g. insurance/payment) • Stock-price manipulation • Brand damage • Reputational damage	• News media manipulation • Damage to economic stability • Damage to the justice system • Damage to the scientific system • Erosion of trust • Damage to democracy • Manipulation of elections • Damage to international relations • Damage to national security

[157] Bingquan Zhu et al., 'Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation,' *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, February 7, 2020, 414–20.

[158] 'Deepmemory,' *Yori Ettema* (blog), accessed May 11, 2021.

[159] Interview Thies

[160] Chesney and Citron, 'Deep Fakes'; Miha Šepec and Melanija Lango, 'Virtual Revenge Pornography as a New Online Threat to Sexual Integrity,' *Balkan Social Science Review* 15 (June 25, 2020): 117–35; Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.'

[161] This table is based on and adapted from: Aengus Collins, 'Forged Authenticity: Governing Deepfake Risks,' 2019. While Collins developed a crosstable differentiating between three 'key potential impacts' (reputational, financial, and manpilation of decision- making) at three different levels (individual, organisational and societal), in this report the types of risks and levels of impact are presented separately: the risks in this table, the impacts in Figure 9. The reasoning is that Collin's potential impact category 'manipulation of decisionmaking' was considered too narrow for the risks this research identified, so a 'societal harm' column fits better. As a consequence, the 'societal level' row became redundant. Furthermore, it was considered that Figure 9 was better able than a table to illustrate the insight that the damage of a deepfake is typically felt at different levels simultaneously, as the impact of a single deepfake often exceeds the individual level.

It is important to note that some of the identified risks relate to harms that have already materialised, others - mainly on the societal level - are entirely plausible with today's technology and are likely to materialise in the future if no measures are taken and deepfake technologies become more readily accessible or broadly used.

5.3. Risk of psychological harms

The creation and publication of a deepfake may cause severe psychological harm to the represented individual. The smearing videos could be used for **bullying, defamation and intimidation**, which could cause profound reputational and psychological damage.

The first applications of deepfake technology arose in a pornographic context, by editing the faces of celebrities into sex videos without their consent. The potential harms of these videos can be similar to *revenge pornography*; a form of cybercrime offence. According to Šepec & Lango (2020) revenge pornography refers to 'non-consensual dissemination of intimate images that were taken with the consent of an individual but with the implicit expectation that these images would remain private'.[162] Whereas anyone could become a victim of revenge pornography, several interviewees stress that non-consensual deepfake pornography has a strong gender dimension as it seems to target almost only women. Individuals portrayed in such videos may also suffer collateral consequences and reputational sabotage or loss of opportunities, for example in the job market. It has also been described as a strategy for silencing speech.[163] Deepfakes may deepen a problematic social phenomena know as social cooling, which means that people avoid seeking public attention because of the risk of becoming a target of deepfakes.

Another important consequence of convincing manipulated videos and audio of people doing or saying things they never did or said is their use for **extortion**. By threatening to expose fabricated content the perpetrators gain power over their victims, for example demanding a fee or following instructions. Thus not just a deepfake itself, but also the use of a deepfake by a malicious actor can cause severe psychological harms. When pornographic images are used for extortion, this is known as **sextortion**.

In addition to psychological harms to the individuals targeted by deepfakes, there are psychological harms to society at large. When people are made aware of the very existence of deepfakes it has been shown to **undermine trust** in visual media.[164] Victims of extortion also describe a general increase of distrust towards others, as it may not always be clear whom the perpetrator is.[165] Even the threat of becoming a future victim of (s)extortion may already cause psychological harm.

5.4. Risk of financial harms

The emergence of deepfakes also gives rise to several risks of financial harms. First, the harms of the above described (s)**extortion** practices may well extend from the psychological domain into the financial domain. Moreover, criminal actions are mostly financially driven. This financial harm may be inflicted on individuals as well as organisations, as employees could be corrupted by extortion. As the

[162] Šepec and Lango, 'Virtual Revenge Pornography as a New Online Threat to Sexual Integrity.'

[163] Sophie Maddocks, ''A Deepfake Porn Plot Intended to Silence Me': Exploring Continuities between Pornographic and 'Political' Deep Fakes,' *Porn Studies* 7, no. 4 (October 1, 2020): 415–23.

[164] Ternovski, Kalla, and Aronow, 'Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments.'

[165] Maddocks, ''A Deepfake Porn Plot Intended to Silence Me.''

creation of deepfake videos can be automated, the process of (s)extortion may also be automated and rapidly scale.[166]

Moreover, deepfake technology may be used to steal identity by attacking biometrics in the verification process for online banking transactions, or of employees in an organisation. This new form of **identity theft** could be used for various goals, such as creating convincing imitations of superiors giving orders or directions to employees. A well-known example is the case of a money transfer of 243.000 British pounds to a Hungarian bank account.[167] Using voice cloning technology, the attacker had pretended to be the CEO of a United Kingdom (UK)-based energy firm and asked the firm's chief to make the transfer. It is also conceivable that criminals could obtain trade secrets, passwords or other important information from organisations in this way, resulting in substantial information security risks and subsequent financial harms.[168]

These types of scams can affect businesses, but also individuals and families. In an evolved version of the 'grandma scam' for instance, criminals are using deepfakes to act as a family member who needs emergency funds.[169]

Deepfakes can also enable numerous other methods of **fraud**. A deepfake video could depict a chief executive inciting hatred, insults or other immoral or illegal behaviour. False statements could also be made about alleged company takeovers or mergers, about financial losses or bankruptcy.[170] When these frauds target publicly traded companies, this may result in **stock market manipulation**. It is conceivable, that even if a company makes a timely clarifying statement, **brand or reputation damage** could still be the consequence, from which the company may not fully recover.

5.5. Risk of societal harms

This risk category is a receptacle of potential adverse impacts of deepfakes in multiple societal sectors and institutions. Vulnerable societal sectors include those that rely heavily on documented evidence, such as insurance, journalism, media and education, and societal and economic systems such as the financial market, the criminal justice system, the political and the science systems. The paragraphs below elaborate on the kinds of harm that could be expected in these contexts.

Manipulation of news media

The risks of deepfakes are often linked to the potential harms of mis- and disinformation[171], recognising the potential to manipulate news media. Deepfake disinformation could for example comprise of attempts to influence public opinion, gather fake campaign donations, and slander public figures. Researchers have demonstrated that a carefully designed deepfake video has a political effect.[172] In their paper, 'Language Models are Few-Shot Learners'[173], the researchers describe in detail the possible harmful effects of their text synthesis system GPT-3. They warn that the 'high-quality text generating capability of GPT-3 can make it difficult to distinguish synthetic text from the human-written text'. The

[166] V Ciancaglini et al., 'Malicious Uses and Abuses of Artificial Intelligence' Trend Micro Research, United Nations Interregional Crime and Justice Research Institute & Europol's European Cybercrime Centre, November 19, 2020.

[167] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.'

[168] Ciancaglini et al., 'Malicious Uses and Abuses of Artificial Intelligence.'

[169] S. S. Sokolov et al., 'Modern Social Engineering Voice Cloning Technologies,' in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2020, 513–16.

[170] Ciancaglini et al., 'Malicious Uses and Abuses of Artificial Intelligence.'

[171] Aya Yadlin-Segal and Yael Oppenheim, 'Whose Dystopia Is It Anyway? Deepfakes and Social Media Regulation,' *Convergence* 27, no. 1 (February 1, 2021): 36–51.

[172] Tom Dobber et al., 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?,' *The International Journal of Press/Politics*, July 25, 2020, 1940161220944364.

[173] Tom B. Brown et al., 'Language Models Are Few-Shot Learners,' *ArXiv:2005.14165 [Cs]*, July 22, 2020.

authors point to several scenarios of misuse. Their list includes misinformation as well as other fraudulent writing. Another risk concerns biases from training data that end up in the models, for example stereotypes and prejudices. It was, for example, found that the models associate words with religious terms that reflect a negative bias towards some religions; for example, words such as 'violent', 'terrorism' and 'terrorist' were associated at a higher rate with Islam than with other religions. The authors believe additional bias prevention measures are necessary to prevent harm. Critics state that the text-synthesis technology community is not investing enough in creating high-quality training sets, based on the false assumption that gathering more training data will always lead to better models. They recommend 'encouraging research directions beyond ever larger language models'.[174]

In addition to such harm to society at large, deepfake disinformation also confronts journalists with the challenge to fulfil their ethical and moral duty to report the truth, which means an increased burden on their part to determine the authenticity of text, audio and graphic materials.

Damage to economic stability

Manipulated news media in turn can damage economic stability. For example, synthetically generated statements about the dispute between Saudi Arabia and Russia regarding oil production quotas could have a negative impact on the price of oil and thus on the global economy. However, the severity of such an impairment of financial markets depends to a great extent on other factors than the quality of the deepfake itself. Bateman concludes there is 'no serious threat to the stability of the global financial system or on national markets in mature healthy economies'.[175] Developed countries would be more likely to be affected in already unstable situations, such as an ongoing economic crisis. In contrast, less developed countries, or rather emerging markets, are exposed to greater danger, as the assumed lack of stabilising institutions make them more susceptible to manipulations.

Damage to the justice and science system

Deepfakes may also damage the justice and science systems. Deepfake videos, voice clones and synthetic texts could be used to create false evidence in criminal court cases or used as evidence for scientific claims. As fraud has plagued science and the courts for years, it has already been criminalised. However, deepfakes may be much harder to detect. Deepfakes therefore raise serious concerns regarding the fundamental 'credibility and admissibility of audio-visual footage as electronic evidence before the courts'.[176] Even when existing validation procedures for audio and video evidence are able to detect deepfakes, the very existence of deepfakes may still influence testimonies, because people may still testify based on what they saw or heard in a deepfake outside of the court.

Erosion of trust

The potential manipulation of news media, science and the justice system leads to a much wider concern of a general erosion of trust in society. It is feared that deepfakes may lead to a situation in which trustworthy information no longer exists.[177] This general loss of trust of people in any kind of information is sometimes referred to as 'information apocalypse' or 'reality apathy'. Galston recently described this looming state of general uncertainty with the following words:

> *'If AI is reaching the point where it will be virtually impossible to detect audio and video representations of people saying things they never said (and even doing things they never did), seeing will no longer be*

[174] Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? &#x1f99c;,' in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 Virtual Event, Canada: Association for Computing Machinery, 2021, 610–23.

[175] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.'

[176] Ciancaglini et al., 'Malicious Uses and Abuses of Artificial Intelligence.'

[177] Regina Rini, 'Deepfakes and the Epistemic Backstop,' *Philosopher's Imprint* 20, no. 24 (2020).

> *believing, and we will have to decide for ourselves - without reliable evidence - whom or what to believe.*'[178]

The conviction that what we see does not reflect the truth, can lead in the final instance to the point that ultimately even truth will not be believed.[179] This effect is also described as the 'liars dividend': those spreading doubt and uncertainty ultimately benefit, because they gain in the ability to mask the truth. The potential use of deepfakes for this purpose means that the technology introduces a new instrument for malicious politicians to gain power at the cost of citizens and journalists.

Damage to democracy

The erosion of trust created by deepfakes is especially disturbing as we live in a time where there is already distress about disinformation campaigns targeting democracies. Deepfakes can be expected to damage democracy in several ways, especially the public debate, elections, the legitimacy of democratic institutions[180] and the power of citizens[181] and politicians. [182] In the following paragraphs these potential problems are described in more detail.

The potential manipulation of news media is problematic as it ties directly into a vital process of democracies: public debate. The integrity and quality of public debate is crucial as it is the main instrument for citizens to formulate their political opinions. However, in order for a public debate to function, there has to be some common sense of reality, which includes a common sense of what the public debate is about, who is participating, and what positions these participants represent. Deepfakes may manipulate all these aspects of the common sense of reality.[183]

There are also debates on the strengthening effect of deepfake technologies on a general change in the culture of the public debate through fragmentation and polarisation of the digital communication. Deepfakes spread by micro-targeting have framing effects on people, who only believe what fits with their own world view; a phenomenon which is also called 'echo chambers'.[184] This could also be used for political manipulation and targeted propaganda. In addition, interviewees indicate this kind of disinformation increases the rise of conspiracy theories.

Deepfakes may also inflict long-lasting damage on the reputation of public figures, including politicians and other elected officials, thereby leading to a **manipulation of elections**. In 2019 for example, a deepfake video was circulating widely in Malaysia that depicts a political aide who seems to admit having had a homosexual relation with a cabinet minister. The video also includes a call for investigating alleged corruption by the minister and led to a destabilisation of the coalition government.[185] The manipulation effect on elections will be most likely if the attacker distributes a deepfake in such a way that there is enough time for it to circulate, but not enough time for the target

[178] William A. Galston, 'Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics,' *Brookings* (blog), January 8, 2020.

[179] Galston.

[180] W Lance Bennett and Steven Livingston, 'The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions,' *European Journal of Communication* 33, no. 2 (April 1, 2018): 122–39.

[181] D. J. Flynn, Brendan Nyhan, and Jason Reifler, 'The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics,' *Political Psychology* 38, no. S1 (2017): 127–50.

[182] S Bradshaw and P Howard, 'Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation' University of Oxford, 2018.

[183] Yiping Xia et al., 'Disinformation, Performed: Self-Presentation of a Russian IRA Account on Twitter,' *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1646–64; Ciancaglini et al., 'Malicious Uses and Abuses of Artificial Intelligence.'

[184] Stephan Lewandowsky et al., 'Technology and Democracy: Understanding the Influence of Online Technologies on Political Behaviour and Decision-Making' JRC, 2020.

[185] Nic Ker, 'Is the Political Aide Viral Sex Video Confession Real or a Deepfake? | Malay Mail,' 2019.

to deflate it . Examples of such disinformation interventions have been found in the elections in the United States in 2016, and in France in 2017.[186]

Damage to national security and international relationships

Deepfakes may also exacerbate social divisions, civil unrest, panic and conflicts, undermine public safety and national security.[187] At the worst, this could cause violent conflicts, attacks on politicians, governance breakdown or threats to international relations. In 2018 for example, a video with Ali Bongo Ondiba, the President of Gabon, was published online. For months before he had not been seen in public and it had become a popular believe that he was in poor health, or even dead.. The video led to a national crisis.[188] A story that the video was a deepfake gained momentum, as it seemed to support a theory that the government was trying to hide the condition of the President. Ultimately, this story led to an unsuccessful coup d'état by the Gabonese military.

These examples show that deepfake videos could likely cause domestic unrest and protests. It is also conceivable that deepfakes even lead to damage to international relationships or international armed conflicts if governments engage in military actions based on false information.[189]

5.6. Cascading impacts

The impact of a single deepfake is not limited to a single type or category of risk, but rather to a combination of cascading impacts at different levels (See Figure 9). First, as deepfakes target individuals, the impact often starts at the individual level. Next, this may lead to harms in a group or organisation. Thirdly, the notion of the existence of deepfakes, a well-targeted deepfake or the accumulative effect of deepfakes may lead to severe harms at societal level.

The infographic below depicts three scenarios that illustrate the potential impacts of one type of deepfake on different levels:

1. **Pornographic manipulated video.** In this scenario a pornographic video is used as a backdrop to blackmail. The potential direct impact is reputational and psychological damage to the person portrayed. The blackmail may extend to the group - for example the family - this person belongs to, or the company this person is associated with. Once pornographic deepfakes become more common, at societal level this category of deepfakes may have an adverse impact on sexual morality.
2. **Manipulated sound clip as evidence.** The latest advancements in audio-graphic deepfakes mean that anyone who has published recordings of their voice could be fabricated into an audio recording that may serve as evidence to make a person look suspicious. On the individual level, getting involved in a court case based on manipulated evidence obviously would have severe consequences. At the organisational level, this means that courts will have to adjust their processes of authenticating evidence or perhaps dismiss audio-graphic evidence completely. This may hamper the course of justice and undermine the functioning of the court system as a whole.
3. **False statement to influence politics.** In a recent study that demonstrated the political effects of a deepfake video, a manipulated statement about religion by a Christian-Democrat

[186] Chesney and Citron, 'Deep Fakes.'

[187] Chesney and Citron.

[188] Sarah Cahlan, 'How Misinformation Helped Spark an Attempted Coup in Gabon,' *Washington Post*, 2020.

[189] Hany Farid, 'Hany Farid: Deepfakes Give New Meaning to the Concept of 'fake News,' and They're Here to Stay,' Text.Article, Fox News, June 16, 2019; K. Hartmann and K. Giles, 'The Next Generation of Cyber-Enabled Information Warfare,' in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300, 2020, 233–50.

politician was used.[190] It shows that this kind of deepfake may lead to reputational damage for the politician as well as a loss of trust in the political party. When such deepfakes are used at large, damage to the public debate is likely, and in the long run, even the position of democratic institutions such as the parliament and the integrity of elections are at stake.

In Chapter 7, these scenarios are further developed in order to capture the regulatory gaps that remain in preventing and addressing these adverse impacts of deepfakes.

[190] Dobber et al., 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?'

Figure 9 - Cascading effects of three types of deepfakes (a pornographic video, manipulated audio evidence and a false political statement) on the individual, organisational and societal level

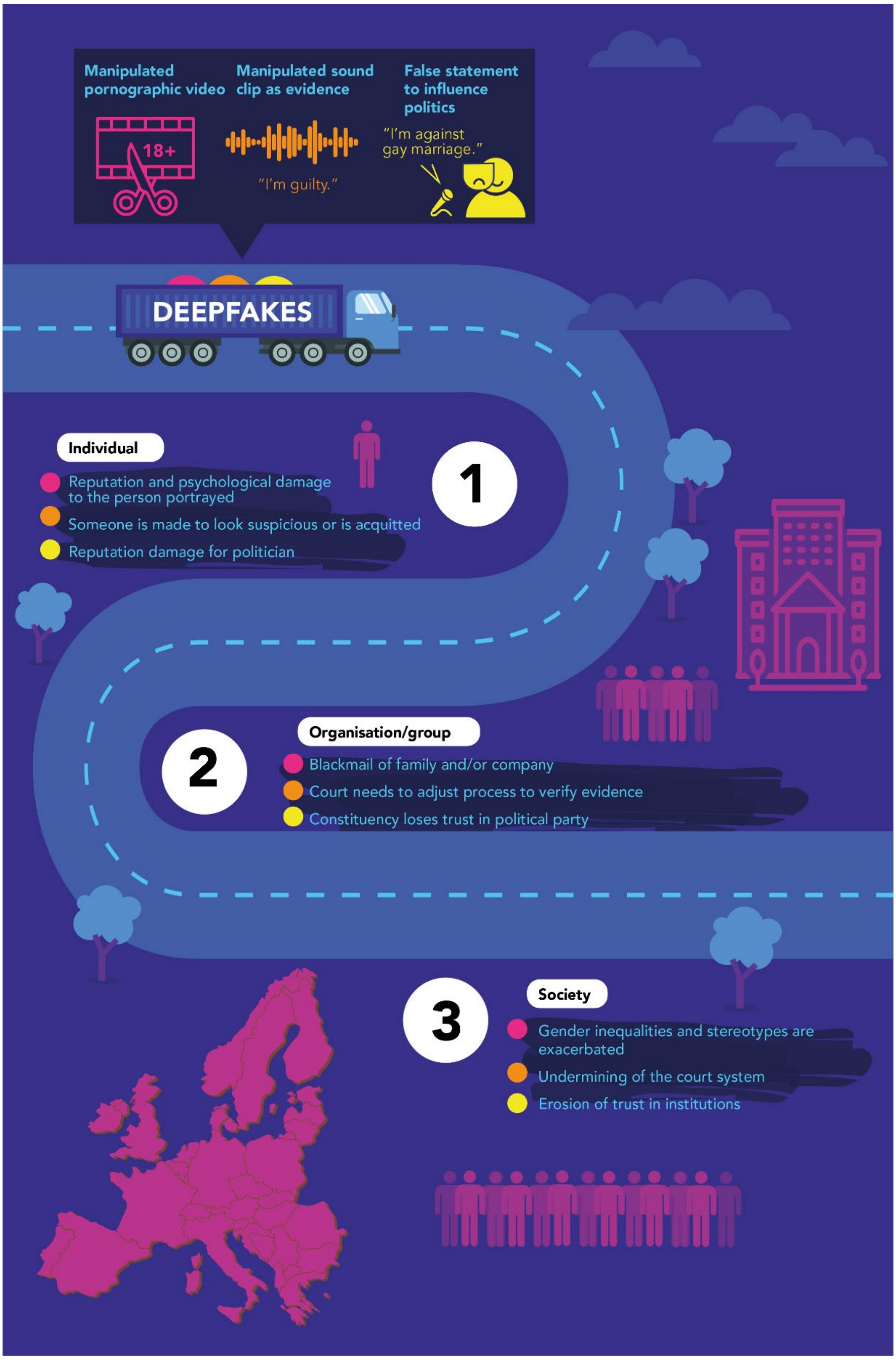

Manipulated pornographic video
18+
Manipulated sound clip as evidence
"I'm guilty."
False statement to influence politics
"I'm against gay marriage."
DEEPFAKES
Individual
● Reputation and psychological damage to the person portrayed
● Someone is made to look suspicious or is acquitted
● Reputation damage for politician
1
2
Organisation/group
● Blackmail of family and/or company
● Court needs to adjust process to verify evidence
● Constituency loses trust in political party
3
Society
● Gender inequalities and stereotypes are exacerbated
● Undermining of the court system
● Erosion of trust in institutions

Image created by Rathenau Instituut.

6. Regulatory landscape

In the previous chapter we saw that there are different types of risks associated with deepfake technologies, and that these risks manifest themselves at different levels. When aiming to mitigate the negative impact of deepfakes, these different levels must be addressed. In this chapter, we assess the current legal basis for protecting individuals, as well as for mitigating the broader societal impact of deepfakes, for example through policies and measures against disinformation.

The regulatory landscape related to deepfakes comprises of a complex web of constitutional norms, as well as hard and soft regulations on both the level of the EU and the Member States. All actors involved in the lifecycle of a deepfake have rights and obligations that need to be taken into account. These actors include the creator of the deepfake, the person(s) depicted in the video; both the victim and the original performer, the author(s) and copyright owner(s) of the original material, the technology developer(s), the intermediary platform that is used for dissemination, and the platform users who upload, view, or share the video.

In this chapter we will only discuss the most relevant legal frameworks in relation to this study. These are:

- AI regulatory framework proposal
- General Data Protection Regulation
- Copyright law
- Image rights
- E-commerce Directive
- Digital services act proposal
- Audio Visual Media Services Directive
- Measures against disinformation
- European Parliament resolutions related to deepfakes

We also summarise the regulatory debate on deepfakes in selected countries outside the EU:

- United States of America
- China
- India
- Taiwan

6.1. AI regulatory framework proposal

Since deepfakes are a product of AI-based technologies, the rules and regulations for the use of AI have important implications here. The European Commission published its proposal for a unified approach to regulating AI in April 2021.[191] The proposal offers various policy options with the aim to enable 'trustworthy and secure applications of AI', while at the same time respecting the values and fundamental rights of EU citizens. To this end, it sets harmonised rules for the development, market placement and the use of AI systems.

[191] European Commission, 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act),' 2021.

The proposed AI regulatory framework takes a risk-based approach to the regulation of AI. The Commission distinguishes between 'unacceptable risk', 'high risk', 'limited risk' and 'minimal risk'.[192] With its proposal, the Commission seeks to ban the use of AI systems that pose an unacceptable risk to the safety and fundamental rights of EU citizens. For AI systems that fall into the high-risk category, providers would be obligated to carry out risk-assessments, provide for documentation and human oversight, and ensure high-quality datasets, among other requirements. For certain AI systems , only minimum requirements are formulated, while applications that represent minimal risk will not be regulated.

It is important to note that the proposal allows for the use of deepfake technologies, but articulates some minimum requirements, most notably regarding transparency obligations. Creators of deepfakes are obliged to label their content so that it should be clear to anyone that they are dealing with manipulated footage. Article 52 (3) provides that 'users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated'.[193] However, Article 52 (3) also provides that this labelling obligation does not apply 'where the use is authorised by law to detect, prevent, investigate prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences'.

In contrast to the use of deepfake technology, the use of deepfake *detection* software by law enforcement authorities falls in the category of high-risk [194], as it could pose a threat to the rights and freedoms of individuals. Detection software is thus only allowed under strict requirements, such as the employment of risk-management systems and appropriate data governance and management practices (see Chapters 2 and 3 of the AI regulatory framework proposal).

While a labelling obligation for deepfakes could be a first step towards mitigating potential negative impacts, the nature and scope of this measure remains unclear. Firstly, the proposal does not include concrete guidelines for such disclosure, leaving open how these should look, and whether providers of deepfake technology should play a role in enabling such labelling. Secondly, the proposal does not include any measures against those users who fail to meet the transparency requirements of Article 52 (3). Article 71 does not specify among the penalties whether and to what extent non-compliance with the requirements of Article 52 (3) is punishable. Lastly, it remains to be seen if malicious actors, who often distribute deepfakes anonymously, will even be affected by these requirements, since they would be able to avoid detection. In Chapter 8 we will make suggestions for improvement of the AI framework in relation to deepfake-related risks.

6.2. General Data Protection Regulation (GDPR)

The creation of a deepfake typically involves the use of personal data. These are data that can be traced back to an individual, or by which an individual can be identified, including for example voice fragments or photos and videos depicting individuals. A deepfake that depicts a natural person can be considered personal data, since it relates to an identified or identifiable natural person.[195] Personal data

[192] Although only 'prohibited AI practices' and 'high-risk AI systems' are explicitly mentioned in the body of the proposal, the Commission introduces the four categories of risk in the 'Q&A New rules for Artificial Intelligence'.

See European Commission, 'New Rules for Artificial Intelligence – Questions and Answers,' Text, 2021.

[193] Exceptions to this obligation under the same Article concern 'the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU'

[194] Point 6(c) of Annex III in conjunction with article 6(2) of the proposed Artificial Intelligence Act.

[195] Article 4 (1) GDPR

may only be processed under certain conditions, since every individual has the right to privacy and data protection. The general rules for processing personal data are laid down in the *General Data Protection Regulation* (GDPR).

It should be clear that 'processing' is a rather broad term, that comprises all possible uses of personal data in the lifecycle of a deepfake. The broad scope of this notion has relevant implications for technology developers and deepfake creators alike, since personal data are not only used to create specific deepfakes, but also to train the software that is used for the creation of a deepfake. Consequently, the GDPR also protects the data of individuals using deepfake technology-based apps such as FaceApp and TikTok. The operation of a service that enables the creation of deepfake videos requires the performance of a Data Protection Impact Assessment (DPIA).[196] The GDPR is thus applicable to the development of deepfake software and applications, and to the creation and dissemination of deepfakes.

The GDPR provides that the processing of personal data always requires a legal basis. There are six possible legal grounds for the processing of personal data[197], but only 'informed consent' and 'legitimate interests' are likely to qualify within the context of deepfakes. When the creator of a deepfake claims to have a legitimate interest for processing someone's personal data, the legitimate interests pursued by the creator may not be overridden by the interests or fundamental rights and freedoms of the person depicted. This could be the case, for example, with an ironic deepfake depicting a famous person. In such a case, the creator could claim the right to freedom of speech for purposes of satire or political commentary.

When legitimate interests are not applicable, the use of personal data for the creation and dissemination of deepfakes needs to be subjected to informed consent by the persons depicted in the video. It is important to note here that consent must be obtained from both the person(s) in the original video and the person(s) who appear in the fabricated video, as the personal data of all of them are processed.[198] If creators of a deepfake fail to obtain prior consent, they are at risk of violating the GDPR. These requirements do not apply to deepfakes depicting deceased individuals, such as historical figures.[199] There are, however, specific laws at the Member State level[200] that require obtaining consent of the heirs before processing the personal data of a deceased person.

The GDPR offers substantial guidance for tackling unlawful deepfake content, and provides victims with the right to correct inaccurate data, or even have it deleted. In every Member State, there is at least one independent supervisory authority responsible for ensuring and enforcing the rules and regulations. Within the context of deepfakes, however, the legal route for victims can be rather challenging. In many cases, it will be impossible for the victim to identify the perpetrator, who often operates anonymously. Moreover, victims might lack the appropriate resources needed for starting a judicial procedure, leaving them vulnerable.

[196] According to Article 35 (1) of the GDPR, the controller needs to carry out a DPIA 'where a type of processing in particular using new technologies [..], is likely to result in high risk to the rights and freedoms of natural persons'. Since providers of deepfake video services often rely on new technologies such as Generative Adversarial Networks (GANs), and collect vast amounts of biometric (face) data which are classified as personal data requiring special protection, these processing activities already seem sufficient to justify the obligation to conduct a DPIA. See also Hewage 2020

[197] Respectively Art. 6 (1) a, Art. 6 (1) b and Art. 6 (1) f

[198] Chaminda Hewage, 'Data Protection in the Wake of Deepfakes,' Infosecurity Magazine, 2020.

[199] Recital 27 GDPR

[200] For example: Ines K. Radmilovic, Tamás Bereczki, and Ádám Liber, 'Hungary Adopts National GDPR Supplementing Legislation,' 2018.

6.3. Copyright law

The creation of a deepfake usually involves the use of existing video- or photographic material, that might be protected by copyright law. Copyright law applies to 'copyright works', and gives the copyright owner the exclusive right to decide what it will be used for. In principle, such works can thus only be used when the author (and thus copyright owner) has given permission. Within the European Union, copyright law continues to be based on national law in Member States, but has become more or less harmonised thanks to EU legislation.

From the list of copyrighted material, photographic works and cinematographic works in particular are most likely to apply within the context of deepfakes. Copyright owners of those works can make claims, and object against the use of their material in a deepfake video. This means that a deepfake creator must in principle always have permission of the copyright owner of the original material, before using the work to create a deepfake.

However, there are restrictions as to what counts as a 'work' that is protected under copyright law. Furthermore the use of copyrighted material to generate deepfakes for scientific use, and purposes of caricature, parody or pastiche is widely permitted under exceptions.

6.4. Image rights

Since individuals generally do not own a copyright interest in their own image, copyright law is not very suitable for individuals to protect their own persona. However, in some EU Member States there are other legal provisions for the protection of a person's image or portrait[201]. Although the protection of image rights in the EU still remains far from harmonised[202], most Member States recognise at least some form of legal protection. Furthermore, in a ruling in 2009, the European Court of Human Rights stated that the right to the protection of one's image is 'one of the essential components of personal development and presupposes the right to control the use of that image'.[203] The contracting states of the *European Convention of Human Rights* (ECHR) should respect this ruling.

The right to the protection of one's image is strongly related to the right to protection of personal life as formulated in Article 8 of the ECHR. According to the Court, a person's image 'constitutes one of the chief attributes of his or her personality, as it reveals the person's unique characteristics and distinguishes the person from his or her peers'.[204] It is deemed essential for the identity of an individual and thus deserves protection. The definition of an 'image' is rather broad and protects not just a portrait, photograph or video depicting an individual, but also 'likeness' or resemblance of a person. Recognition could be sufficient for image rights to come into play.

This implies that, in jurisdictions where image rights are protected, the use of an image for the creation of a deepfake could be unlawful. However, the right to the protection of one's image is not an absolute right, which means that fundamental rights and freedoms of others should always be taken into account. In addition to the creator's rights (freedom of speech, parody, political commentary, etc.), the context in which the image is used is also relevant for deciding if the use of an image is lawful or not.

[201] E.g. France, Germany, The Netherlands and Spain

[202] Tatiana Synodinou, 'Image Right and Copyright Law in Europe: Divergences and Convergences,' *Laws* 3, no. 2 (2014): 181–207.

[203] 'Right to the Protection of One's Image,' *European Court of Human Rights*, 2020.

[204] 'Right to the Protection of One's Image.'

6.5. e-Commerce Directive and the digital services act

6.5.1. e-Commerce Directive

A basic building block for regulating online content is the EU Directive on electronic commerce (e-Commerce Directive),[205] which was adopted in the early days of commercial use of the internet. The directive decides that intermediary service providers are not subject to any obligation to monitor the information flowing through their channels.[206] The aim of this regulation was to facilitate the economic boom of service providers by refraining from imposing excessive regulatory requirements. The Directive provides that no ex ante verification of content will be carried out. Nevertheless, the directive also provided for the obligation that providers must remove content as soon as they become aware of the existence of illegal content.

In principle, the e-Commerce Directive already enables the removal of illegal deepfake content. However, it does not contain a clear definition of what exactly is meant by illegal content, which makes it unclear what distributors have to comply with. Furthermore, the Directive harmonised the conditions for *releasing* providers from liability, but not the conditions that must be met in order to *establish* liability. The Commission recognised as early as 2012 that harmonisation in this area was insufficient, but until recently it refrained from regulatory measures and focused on encouraging self-regulation by platforms.

6.5.2. Digital services act

Against the backdrop of emerging legal fragmentation,[207] the Commission announced EU-wide harmonisation of liability rules and content moderation obligations for digital platforms, services and products. In late 2020, it presented a proposal for a legislative package comprising of the *digital services act* (DSA) and the *digital markets act* (DMA). This package will replace the E-commerce Directive.[208]

Since the DSA applies to content on social media platforms, it is of relevance for the dissemination of deepfakes. The DSA stipulates that intermediary providers which moderate user-generated content must make transparent which moderation rules apply, and which measures they implement to enforce these. It also stipulates that platforms using a notice-and-takedown-procedure must create a system by means of which illegal content can be reported. Platforms above a certain size must implement a procedure by which those affected can appeal against any blocking.

Finally, the law will require all providers to provide more transparency about any blocking that has taken place. The database must contain information about the reason for blocking and the respective

[205] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')

[206] This is reflected in particular in Articles 12 to 14

[207] Against the backdrop of the Commission's initial refusal to adopt hard regulatory measures, some Member States forged ahead in the area of platform regulation (Madiega, 2020, p. 8). In this context two developments in particular stand out: the German federal government's NetzDG and the French Loi n° 2018-1202. After many years of voluntary commitments to improve law enforcement on the part of providers did not lead to any noticeable improvements, the adoption of the NetzDG was aimed at being able to enforce applicable law, especially against operators of very large social networks. The French Loi n° 2018-1202 was created for similar reasons, but with a focus on preventing fake news from influencing elections. Differences exist, for example, in the defined degree of illegality and the actors who are held accountable. For example, the German law introduces a distinction between manifestly illegal content, which must be deleted immediately, and illegal content, which must be decided upon within seven days. In both cases, the platform operators are to decide for themselves. The French law, on the other hand, only provides for the immediate deletion of patently illegal content following judicial authorisation. There are greater similarities between the two regulations in terms of procedural and transparency provisions (Madiega, 2019, p. 11 f.; Pollicino et al., 2020, 25).

[208] European Commission, 'The digital services act Package,' 2021.

complainant. Unfortunately, the fundamental challenge of how to classify different forms of content as illegal or non-illegal remains unresolved in the DSA.

6.6. Audio Visual Media Services Directive

The EU Audiovisual Media Services Directive (AVMSD) was revised and adopted in 2018 in response to the extension of the media landscape with online video-sharing platforms. The directive legally defines video-sharing platforms, thereby creating the opportunity for Member States to create regulations directed specifically at these services.

The AVMSD contains several guidelines on preventing harm, especially drawing attention to the protection of the wellbeing of minors. Member States are directed to regulate video-sharing services in order to prevent impairment of the physical, mental or moral development of minors and to offer effective parental controls. Pornography and violent content should be treated by the strictest measures, such as age-verification, PIN-codes, clear labelling or automatic filtering. The AVMSD calls for regulations that require video-sharing platforms to detect the nature of the content shared and implement measures in the interest of the viewer, creator and general public. The AVMSD thus contains provisions to respond to, for example, the distribution of non-consensual pornographic deepfakes.The Directive recognises that Member States will have to balance the regulation of harmful content with applicable fundamental rights, such as the freedom of expression and respect for private life.

6.7. Measures against disinformation

Since deepfakes can be used as a vehicle for disinformation, the legal framework related to disinformation is also relevant in this context. The discussion around tackling disinformation in the EU commenced after the start of the Russian war against Ukraine in 2014. Soon, the first initiatives were launched to counter disinformation on the internet.[209] These measures were taken from a predominantly foreign policy and security perspective.

6.7.1. Code of Practice on Disinformation

These steps initially led to the European approach on tackling online disinformation with the publication of the Code of Practice on Disinformation in 2018. In addition to measures such as closing fake accounts or preventing bot-driven activity, a key component of the Code is an attempt to tame online political advertising. This is to be achieved by platform operators making a distinction between political and non-political content and demonetising political advertising that contains disinformation. The Code invited platform operators to voluntarily submit to a number of best practices in order to achieve greater transparency and accountability, especially via the periodic publication of reports. The Code was signed by Mozilla, Twitter, Facebook, Google and some other stakeholders in October 2018. Microsoft and TikTok joined later in 2019 and 2020, respectively.[210]

Several studies, however, indicated that the Code was lacking a meaningful possibility to measure its effectiveness, and that the published transparency reports often did not contain important information.[211] As a result, the effectiveness and efficiency of the Code itself was called into question.[212]

[209] These initiatives included: the call for action on the European Council of 19th and 20th of March 2015, the European Parliament's resolution of 2017 on online platforms and the digital single market, which urged the Commission to adopt hard regulatory means to deal with fake news, i.e. the revision of the e-Commerce-Directive, a public consultation of the Commission on fake news in late 2017 and the appointment of a High-Level Expert Group on Fake News and Online Disinformation in early 2018.

[210] European Commission, 'Code of Practice on Disinformation,' 2021.

[211] 'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice' ERGA, 2020.

[212] Iva Plasilova et al., 'Study for the Assessment of the Implementation of the Code of Practice on Disinformation,' 2020.

In its own evaluation the European Commission recognised several shortcomings, and has identified the need for improvements, pointing at the digital services act as an important opportunity to realise those improvements.[213]

6.7.2. EU action plan against disinformation

While the Code was dedicated to strengthening the accountability of private social media operators, work was also done on further measures to counter disinformation. Against the backdrop of perceived influence by foreign actors in both the Brexit referendum in 2017 and the U.S. presidential election in 2016, the EU action plan against disinformation was published in 2018, with the aim of protecting the upcoming European elections from similar influence.

Four sets of measures were proposed in the action plan: (1) improving the capabilities of Union institutions to detect, analyse and expose disinformation; (2) strengthening coordinated and joint responses to disinformation; (3) mobilising the private sector to tackle disinformation; (4) raising awareness and improving societal resilience.[214] As a result, a number of specific measures were adopted as part of the so-called 'Election package'.[215] A Report on the implementation of the *Action Plan on Disinformation* by the High Representative of the Union for Foreign Affairs and Security Policy and the Commission noted that 'it contributed to expose disinformation attempts and to preserve the integrity of the elections […] while preserving freedom of expression.'[216] The Commission made similarly positive comments on the implementation of the measures in a communication from late 2019.[217] However, several research institutes have judged these measures to be inadequate with regard to the danger posed by deepfakes.[218]

6.7.3. European democracy action plan

The most recent measure in this area is the European democracy action plan, which was unveiled in December 2020. Under the three pillars: 1. Promote free and fair elections; 2. Strengthen media freedom and pluralism; 3. Counter disinformation, there is a wide range of proposed measures and concrete announcements. The most relevant announcement of the European democracy action plan with regard to the dimension of disinformation refers to the transition from self-regulation to co-regulation, in response to the insufficient implementation of the Code of Practice Against Disinformation. To this end, it is envisaged that the existing Code will be transformed into a co-regulatory framework in line with the DSA to strengthen the accountability of platform operators and to better achieve the objectives already spelled out in the Code. It is therefore planned that the Commission will issue guidance on how to enhance the Code in spring 2021, which will serve an expanded group of stakeholders (now consisting not only of platform operators, but also advertisers, representatives of traditional media, civil society actors, fact-checkers and academics) as a basis for discussion on the revision of the Code. The updated Code should make it possible to better assess the trustworthiness of information, while at the same time increasing the visibility of accurate information in cooperation with scientists and fact-checkers.[219]

[213] 'Assessment of the Code of Practice on Disinformation – Achievements and Areas for Further Improvement' European Commission, 2020.

[214] European Commission, 'Action Plan on Disinformation: Commission Contribution to the European Council,' 2018.

[215] See documents: COM (2018), 637; C (2018) 5949; COM (2018) 638.

[216] High Representative of the Union for Foreign Affairs and Security Policy, 'Report on the Implementation of the Action Plan Against Disinformation' Brussels, 2019.

[217] European Commission, 'Security Union: European Commission Presents the Twentieth Progress Report,' Migration and Home Affairs - European Commission, October 30, 2019.

[218] Sarah Bressan, 'Can the EU Prevent Deepfakes From Threatening Peace?,' Carnegie Europe, 2019.

[219] European Commission, 'European Democracy Action Plan,' 2020.

6.8. European Parliament resolutions related to deepfakes

The European Parliament has been consistently involved in EU-wide activities aiming to protect democratic elections against manipulative interventions and disinformation. Additionally, it has taken concrete action to deal with the undesirable consequences of artificial intelligence by means of several resolutions and reports. Here, the most relevant pieces are summarised.

In its 2017 resolution, the Parliament urged the Commission to adopt hard regulatory means to deal with fake news, i.e. the revision of the e-Commerce-Directive.[220] The May 2018 parliamentary resolution on media pluralism took up additional proposals that were not related to deepfakes, but are highly relevant to the discussion of deepfakes: full transparency in the use of algorithms, artificial intelligence and automated decision-making with regard to the arbitrary blocking, filtering and removal of internet content (No 25); the importance of independent and impartial certified third-party fact-checking organisations (Nos 32 and 33); obligations and instruments in relation to source verification (No 32); the enabling of users to report and flag potential disinformation (No 33); and the displaying and labelling of disinformation revealed as such to stimulate public debate and prevent re-emergence of the content (No 33).[221]

Specific mention of deepfakes can be found in various parliamentary positions, such as the Parliament's February 2019 resolution calling on the Commission to introduce a labelling requirement for producers of deepfake material or synthetic videos.[222] The proposal to introduce a labelling requirement is reflected in a number of parliamentary documents.[223] The call for the introduction of strict limits (or other protective measures such as thorough investigations into hostile campaigns) on the use of deepfakes in the context of elections can be found in almost all of these documents. [224] In a 2020 report on the intellectual property rights for the development of artificial intelligence, the Parliament calls for increased awareness-raising and media literacy, in order to combat the possibility of mass manipulation through deepfakes.[225]

The most comprehensive and recent document with regard to the discussion of the deepfakes issue is the resolution of 19 May 2021, on 'Artificial intelligence in education, culture and the audiovisual sector'. In addition to the proposals already mentioned above, this resolution contains various forward-looking proposals. These include the importance of raising awareness of the risks of deepfakes and improving digital literacy (No 90); addressing the increasing difficulty of detecting and labelling false and manipulated content by technological means (No 91); calling upon the Commission to introduce appropriate legal frameworks to govern the creation, production or distribution of deepfakes for malicious purposes (No 91); the promotion of the further development of detection capabilities

[220] EP. „Resolution on online platforms and the digital single market'. P8_TA(2017)0272, June 15th 2017.

[221] EP. „Media pluralism and media freedom in the European Union European Parliament resolution of 3 May 2018 on media pluralism and media freedom in the European Union', P8_TA(2018)0204, May 3th 2018.

[222] EP. „Resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics'. P8_TA(2019)0081, February 12th 2019, Nr. 178.

[223] See for example: LIBE Committee. „Opinion of the Committee on Civil Liberties, Justice and Home Affairs for the Committee on Legal Affairs with recommendations to the Commission on the framework of ethical aspects of artificial intelligence, robotics and related technologies'. PE652.296v02-00, September 22nd 2020.

[224] EP. „Recommendation of 13 March 2019 to the Council and the Vice- President of the Commission / High Representative of the Union for Foreign Affairs and Security Policy concerning taking stock of the follow-up taken by the EEAS two years after the EP report on EU strategic communication to counteract propaganda against it by third parties'. P8_TA(2019)0187, March 13th 2019.

JURI. „Report on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice'. PE653.860v02-00, January 4th 2021.

[225] EP. „Report on intellectual property rights for the development of artificial intelligence technologies.' P9_TA (2020)0277 October 20th 2020

(No 92); as well as improving transparency with regard to what content is displayed to platform users and giving them greater freedom to decide whether and what information they want to receive (No 93).[226]

6.9. Regulatory debates in selected countries

The debate to mitigate the harmful effects of deepfakes through regulation is not yet well advanced in most countries. In the following, we provide an overview of activities and measures from selected countries *outside* the EU from which the EU institutions could learn.

6.9.1. United States

In the United States, legislation was first enacted in a few states. The first measures at the federal level have only recently been added. California and Texas were the first states to pass laws in 2019. Under the new California law (AB 730), it is illegal to distribute manipulated content featuring political candidates within a 60-day period before an election that is intended to injure the candidate's reputation or to deceive a voter into voting for or against the candidate. The Texas law is very similar to California's, but only prohibits distribution within a 30-day period. Both laws drew considerable criticism, particularly questioning their compatibility with the right to free speech.[227]

The laws of some states, which are primarily concerned with protection against pornographic deepfakes, were assessed less critically.[228] For example, a law passed in Virginia in July 2019 criminalises the dissemination of such content if it is intended to coerce, harass or intimidate a person. Another California law (AB 602) introduced a private right of action for individuals seen in pornographic deepfakes to simplify the individual complaint process. A New York law passed in late 2020 also introduces the right of private action against pornographic deepfakes, but includes an additional element. This is because it establishes the post-mortem right of publicity to protect an artist's likeness - and hence possible deepfakes of that person - from unauthorised commercial exploitation for 40 years after his or her death.[229]

The measures already adopted at the federal level are limited to the systematisation and institutionalisation of the collection of information, so that further informed action could follow later based on the information made possible by these laws. The *Identifying Outputs of Generative Adversarial Networks Act* (IOGAN Act) provides for the National Science Foundation (NSF) to support research into the production and authenticity analyses of deepfakes, for the National Institute of Standards and Technology (NIST) to conduct research on deepfake standards, and for both institutions to work jointly and in cooperation with the private sector on ways to detect Deepfakes. In addition, for the second year in a row, the U.S. National Defense Authorization Act (NDAA) also includes measures intended to address deepfakes. The 2021 NDAA directs the Department of Homeland Security (DHS) to produce an annual report on 'digital content forgeries' for five years. Unlike the 2020 NDAA, which focused only on reports on the use of deepfakes by foreign states, the DHS has now also been directed to broaden its perspective to look not only at how foreign governments use deepfakes, but also at the broad range of threats that deepfakes pose to the public. In addition, the act also requires DHS to research ways to generate, detect, and counter deepfakes.[230]

[226] EP. „Resolution of 19 May 2021 on artificial intelligence in education, culture and the audiovisual sector'. P9_TA(2021)0238 May 19th 2021.

[227] Matthew Feeney, 'Deepfake Laws Risk Creating More Problems Than They Solve,' Regulatory Transparency Project, 2021.

[228] Ibid.

[229] Matthew Ferraro and Louis Tompros, 'New York's Right to Publicity and Deepfakes Law Breaks New Ground,' 2020.

[230] Scott Briscoe, 'U.S. Laws Address Deepfakes,' 2021.

Other measures at the federal level have either already failed or are still under political discussion. One discussion relevant to the EU revolved around the proposal for the 'deepfake accountability act',[231] which was put on the table in 2019. The proposal stipulates that deepfakes must be labelled as such and that otherwise creators will face heavy penalties. The proposal has been criticised for not curbing the actual goal of preventing the distribution of harmful deepfake videos. After all, malicious deepfake creators could remain undetected by using advanced technologies and would therefore not change their behaviour, while at the same time the creators of legitimate videos would be subjected to unnecessary burdens.[232]

Some of the measures also include elements for making platforms liable - for example, ways of amending Section 230 of the Communications Decency Act, which provides for exemption from liability for content providers, similar to the e-Commerce Directive. But these debates are always conducted against the very absolute US understanding of free speech, on the one hand, and are limited in that they are not intended to impose too great a regulatory burden on platforms, on the other. As the ongoing discussions around the DSA show, the EU is much more concerned with striking a balance between conflicting rights and interests in this area.

6.9.2. India

India does not yet have specific laws regulating deepfakes. In addition to data protection and copyright laws, there is also discussion about tightening the law to encourage social media platforms to take a tougher stance against deepfakes and to hold them liable, if they don not.[233]

A particular feature of the Indian debate is the focus on the fact that the existing laws do not cover the possibility of creating deceased people's deepfakes. While the comparable New York law described above embeds the protection of the image of the deceased in a right of publicity, in India this protection is discussed in the context of data protection law. In this context, reference is made to supplementing data protection laws so that the survivors of deceased persons can watch over their personal data, as is the case in the national data protection laws of Hungary or Spain.[234]

6.9.3. China

Deepfakes have already become a widespread cultural phenomenon in China. Various apps for generating deepfakes have enjoyed massive success in recent years.[235] Like other governments, the Chinese government is striving to curb potential harmful effects emanating from deepfakes. To this end, the government passed a law that came into effect on January 1, 2020. The law stipulates that all deepfake videos or audio content, or content created using deep learning algorithms or VR technologies must be labelled accordingly by the app providers. The law obliges platform operators to independently identify and mark or remove unlabelled content. According to the new law, the production and spreading of fake news is forbidden and must therefore be deleted immediately upon identification.[236]

[231] Yvette D. Clarke, 'Text - H.R.3230 - 116th Congress (2019-2020): Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019,' legislation, June 28, 2019.

[232] Zachary Schapiro, 'DEEP FAKES Accountability Act: Overbroad and Ineffective – Intellectual Property and Technology Forum,' 2020.

[233] In India, for example, the debate revolves around the Information Technology Act 2000, which is the equivalent of the European e-commerce directive. See: Purvi Nema, 'Are Indian Laws Equipped To Deal With Deepfakes?,' *The Journal of Indian Law and Society Blog* (blog), July 19, 2020.

[234] Simran Jain and Piyush Jha, 'Deepfakes in India: Regulation and Privacy,' *South Asia@LSE* (blog), May 21, 2020.

[235] Zehi Yang, 'Chinese Deepfakes Are Going Viral, and Beijing Is Freaking Out,' Protocol, March 19, 2021.

[236] Karen Chiu, 'China Announces New Rules to Tackle Deepfake Videos,' South China Morning Post, November 30, 2019.

The responsibility to enforce the rules was given to the Cyberspace Administration of China (CAC). Because Chinese authorities also face the challenge that, despite the ban, criminals or members of the opposition could continue to distribute content that is illegal or politically unwelcome, the law includes complementary measures to enable effective enforcement of the following rules:

- Users must register on platforms with identifiable information such as government IDs or cell phone numbers, in accordance with the *Cybersecurity Act*.
- Platforms should establish easy-to-use complaint channels.
- Audio and visual services should issue industry standards and guidelines and establish a credit system.
- Government departments must organise regular inspections to ensure that platforms regulate online audio and video in accordance with service agreements.[237]

Most recently, it was reported in mid-March 2021 that the CAC had convened a meeting with 11 Chinese platform operators, at which the demands of the regulator and the Chinese government were once again emphasised.[238]

6.9.4. Taiwan

While the state of Taiwan does not yet have specific legislation to mitigate the harmful effects of deepfakes, it is worth noting the country's strategy to deal with fake news, which may also provide valuable insights for the EU debate.

Due to China's ambitions to annex Taiwan, Taiwan has been exposed to a massive flow of disinformation from the neighbouring country for quite some time. Taiwan's so-called 'nerd immunity' strategy relies on the deployment of hundreds of professional fact-checkers. Taiwan effectively leverages the historically high level of engagement of civic society with government. In addition to the fact-checkers, the country is making a special effort to train the general population to recognise false news and to involve them in checking the truth of online content. Unlike existing strategies in the EU, the key point of this idea is that citizens are not only trained to recognise content, but also to take active action against its dissemination, for example by disseminating corrective content in a creative (and witty) way.[239]

[237] Lavender Au, 'China Targets 'deepfake' Content with New Regulation · TechNode,' TechNode, December 3, 2019.

[238] 'China Regulators Held Talks with Alibaba, Tencent, Nine Others on 'deepfake' Tech,' *Reuters*, March 18, 2021.

[239] Nicola Smith, 'Taiwan Builds 'nerd Immunity' to Resist Chinese Disinformation Campaigns,' *The Telegraph*, June 13, 2020.

7. Regulatory gaps

The previous chapter shows that some regulation already applies to deepfakes, but questions remain about how this works in practice, and to what extent existing regulation protects the rights of victims and deters or punishes perpetrators. We analyse persisting regulatory gaps by developing three scenarios based on the examples provided in Figure 13: a deepfake pornographic video, a false political statement, and manipulated audio evidence.

7.1. Deepfake pornography

Setting the stage

Imagine that you are a female investigative journalist who frequently writes critical commentary on political and socio-economic issues in your country. Recently, you wrote a piece on a corruption scandal that involved several politicians. You are used to being stifled by adherents of the politicians you write about, who often post cruel gossip and lies about you on social media. One day, you receive a message from a friend saying that there is a pornographic video of you circulating online. You are certain that this is impossible, but when you see the video, you are shocked. It is your face copied on someone else's body. It's a deepfake video, intended to harm and discredit you and your work.

Even though this is a fictional scenario, situations like these are already happening in real life.[240] The vast majority of deepfake videos that are currently online consist of non-consensual deepfake pornography.[241] There are various ways in which these videos can be used for harmful exploitation. Citron and Chesney argue that deepfakes can be used for extortion and sabotage, for example when victims are forced to provide money, secret company information, or explicit material (a practice known as 'sextortion') to prevent the release of deepfakes.[242] In the example above, however, the video is already released, merely with the intention of reputational sabotage. What can be done to prevent such harmful practices? Does the current legal framework suffice to protect victims of deepfake pornography? In this short case study, we will assess the impact of deepfakes like these, as well as the question of how the regulatory framework in place addresses these impacts.

Actors involved

Before assessing the potential impact and legal context surrounding pornographic deepfakes, we should make clear which actors play a role. Typically, there is someone that creates the video (**the perpetrator**), and someone that is depicted in the video (**the victim**). But there are many more actors involved in the so called 'lifecycle' of deepfakes. We can think, for example, of the company that developed the software or that was used to create the deepfake (**the technology provider**). For the creation of deepfake pornography, the creator typically uses existing pornographic material, and superimposes the face of the victim onto the body of the persons whose bodies are still being represented in the deepfake (**the original performers**). Additionally, the company that produced the original film content **(the original author)** also plays a role in the legal context. Furthermore, there is the online forum or any kind of intermediary service that is being used for the dissemination of the

[240] In 2018, Indian journalist Rana Ayyub became the victim of a deepfake pornography video. According to Ayyub, the video was intended to silence her. See: Rana Ayyub, 'I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me,' HuffPost UK, November 21, 2018.

[241] Patrini, 'Mapping the Deepfake Landscape.'

[242] Chesney and Citron, 'Deep Fakes.'

video (**the platform**), and there are people (**the platform users**) who upload, view or share the video on that particular website.

Potential impact

The impact of deepfake pornography on someone's private life can be very severe, including jeopardising their personal integrity, reputation and safety. Deepfake porn is, at least today, arguably the most shocking variant of deepfake videos. The depiction of explicit content without the victim's consent can be humiliating and demeaning. Schick argues that 'whatever its guise, the exposure, humiliation and fear that come along with being targeted in this way are devastating its victims'.[243] Citron and Chesney argue that scripting an individual into fake porn not only undermines their personal integrity and human agency, but also 'reduces them to sexual objects, engenders feeling of embarrassment and shame, and inflicts reputational harm that can devastate careers'.[244] This could result in serious physical and mental harm.

However, the example also illustrates how the impact of a single deepfake is not restricted to the person that is being addressed by it. There could be negative impacts for the family of the victim, who, just like her, might feel humiliation and fear for her safety. Additionally, harm to the professional reputation of a journalist could also have negative consequences for the company she works with. If readers of the newspaper question her integrity, they might not want to buy that newspaper anymore.

The scenario also makes clear how the impact of deepfakes can exceed the personal and group level, and also has broader societal implications. According to research, over 95% of the applications of deepfake technologies were used for pornography.[245] Our scenario illustrates that the impact of deepfake pornography is highly gendered, in that the technology is mainly used to attack and discredit women. Chesney and Citron argue that 'in all likelihood, the majority of victims of fake sex videos will be female'.[246] Kalf argues that deepfake pornography can be understood as a form of sexual violence that can be used as a tool to maintain or deepen existing gender inequalities.[247] Similar to actual pornography, deepfake pornography portrays an unrealistic image of women as merely sex objects, affecting their societal position.

Moreover, the example makes clear that deepfakes can form a threat to the functioning of both our information ecosystem and democracy. Deepfakes can be used for the deliberate misconstruction of truth, and the spread of fake news with the aim of causing dismay. However, they can also be used to discredit and eliminate certain individuals in the public debate, like the investigative journalist in our scenario. Journalists play an important role in democratic societies. They serve as a watchdog for promoting democratic accountability and transparency of politicians. In our scenario, the video is aimed to harm the integrity of the journalist, and discredit her work. But this could be done to anyone that plays a role in politics or public debate. When deepfakes are targeted at journalists, politicians, judges or other public figures, with the aim of discrediting or blackmailing them, they thus also represent a danger to free speech and impact the functioning of democracy as a whole.

[243] Schick, *Deep Fakes and the Infocalypse*.

[244] Chesney and Citron, 'Deep Fakes.'

[245] Patrini, 'Mapping the Deepfake Landscape.'

[246] Chesney and Citron, 'Deep Fakes.'

[247] Sanne Kalf, 'What Does a Feminist Approach to Deepfake Pornography Look Like?,' October 24, 2019.

Legal context

In the actor analysis, we have seen that several actors play a role in the lifecycle of a pornographic deepfake. This means that we should take all fundamental rights into account that they could potentially claim, and might be competing. To start, the deepfake video in our example is the result of an unauthorised exploitation and reengineering of other people's images and thus, their personal data. This means that the investigative journalist (the victim) depicted in the video could claim image rights and data protection rights. The author of the original images could claim copyright with regard to the modification of the original film content, while the original performers could also make claims with regard to image rights.

In some deepfake cases, creators of deepfakes can claim that their deepfake is lawful and legitimate, for instance when it provides entertainment, satire, or social or political commentary. These are fundamental rights. But with regard to a non-consensual deepfake pornography video, as in our scenario, it is hard to think of any lawful and legitimate purpose. It is therefore unlikely that the perpetrator could successfully make claims based on these rights. While in the EU, a specific legislative intervention or criminalisation of deepfake pornography is still lacking, this does not mean that the law does not provide any guidance at this point.

Kirchengast argues courts will 'most likely deal with harms accrued by deepfake production and distribution through known categories of criminal, civil and administrative law'.[248] In most European countries, there are provisions that can be used when dealing with non-consensual deepfake pornography. Even if these provisions do not explicitly mention deepfake pornography, they can be useful for the victim to claim rights. The current laws already provide some guidance here.

The enforcement of the law, however, is still rather challenging. As it stands today, only the perpetrator is liable. However, many perpetrators go to great lengths to initiate such attacks at such an anonymous level that neither law enforcement nor platforms can identify them. However, it should be taken into account that the negative effect of the deepfake porn attack described above is particularly strong when the service used by the attacker to produce the video enables particularly authentic videos, and when many people distribute the video while the platforms do not prevent this. The extent to which these other actors can or should also be held accountable, is a difficult question.[249]

Even though their rights are protected under law, victims of deepfake pornography are often not in a strong position to do anything about it. At present, the legal roadmap for victims of deepfake pornography often remains unclear. At the point where a deepfake is circulating on the internet, the individual typically loses control over the video. Kirchengast concludes that 'once an image reaches social media, it may not be able to be removed or deleted (..) the (non-legal) burden often rests with the social media user flagging the image and making a case for removal'. It is also questionable whether existing law, which places the responsibility for holding offenders liable on the victim, is appropriate in such cases. After all, the dissemination of such a video already causes considerable psychological damage to the victims, so that they are hardly in a position to take appropriate countermeasures.

[248] Tyrone Kirchengast, 'Deepfakes and Image Manipulation: Criminalisation and Control,' *Information & Communications Technology Law* 29, no. 3 (September 1, 2020): 308–23.

[249] Edvinas Meskys et al., 'Regulating Deep Fakes: Legal and Ethical Considerations,' *Journal of Intellectual Property Law & Practice* 15, no. 1 (January 1, 2020): 24–31.

7.2. False political statement

Setting the stage

Political propaganda and fake news for political gain are by no means new phenomena. Disinformation, like lies and false insinuations against political opponents, has been around for centuries.[250] But why say something *about* someone, if you can make them say it themselves? Enter deepfakes. Deepfakes offer malicious actors new opportunities for manipulating public opinion. Imagine a situation in which organised actors are aiming to undermine trust in European politics. To this end, they produce a deepfake video that shows several European health ministers in a confidential conversation, saying that they are deliberately withholding vaccine supplies. They distribute the deepfake via different platforms using 'social bots'-algorithms that autonomously produce content and imitate human behaviour.[251] The video spreads quickly as it gets picked up by other social media users who believe the video is authentic and redistribute it.

Until now, deepfakes like these have not yet caused great stir. But there are already some quite convincing examples of manipulated videos of politicians. Think of the one that showed Barack Obama calling Donald Trump a 'complete dipshit'[252], or the video of Nancy Pelosi, that was slowed down to make her appear drunk.[253] As the technology improves, the possibility for a successful deepfake attack on politicians becomes more conceivable. In this scenario study, we will assess the potential impact of the deepfake example described above, as well as its current legal context.

Actors involved

Before assessing the potential impact and legal context of this case scenario, we should make clear which actors play a role in it. In the first place, there is an organised actor that created the video, in our case a foreign intelligence service (**the perpetrator**), and the European health ministers who are depicted it (**the victims**). But there are other actors that play a role. This would include the company that developed the software or service that was used to create the deepfake (**the technology provider**). But unlike with other sorts of deepfakes, this video was *not* created by superimposing someone's face on another person's body. The video material is authentic, but the voice and the corresponding facial features were manipulated. Therefore there are no third party actors involved, such as original performers or authors. The video was distributed via various online fora and intermediary services (**the platform**) using various 'social bots'. Lastly, there are the people who receive the video, might believe the video is true and maybe redistribute it via their accounts (**the platform users**).

With regard to the dissemination of the video, it is relevant to mention that the deepfake was posted by 'social bots'- fake accounts. This makes it harder to identify the source of the deepfake. The bots would post the deepfake in a timely manner, but with sufficiently large time intervals so that technical detection based on analysis of the temporal overlap of the posts would fail. By giving each bot an individual 'personality', the posts that are dropped would match the bot character, making them seem more credible both to the real people viewing them and to bot detection methods.[254]

[250] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy.'

[251] 'How Powerful Are Social Bots?' Günther Thiele Foundation, 2018.

[252] Buzzfeed, 'You Won't Believe What Obama Says in This Video ,' Twitter, 2018.

[253] Washington Post, *Pelosi Videos Manipulated to Make Her Appear Drunk Are Being Shared on Social Media*, 2019.

[254] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.'

Potential impact

In the first place, a deepfake depicting politicians could have a severe impact on both their professional and private lives. A video showing European health ministers stating that they are deliberately withholding vaccine supplies, could mean serious harm to their professional integrity and reputation, and thus to their careers. Moreover, it could even put them in danger, when disenchanted citizens take the deepfake to be true, and turn angry. This means that there are various harmful outcomes possible at the personal level. But unlike with, for example, deepfake pornography, the attackers are aiming for more than just reputational damage.

At the organisational level, we can think of the effect that the deepfake would have on the European Union and the political ecosystem in which different countries work together. The attack could lead to geopolitical tensions between the European Union and the foreign state from which the attack was launched. Suppose that domestic and allied intelligence services manage to find out the operating location of some of the accounts, and conclude that the attack was carried out from a foreign state, this would lead to political tensions.

At the societal level, deepfakes like these can have severe implications. They provide malicious actors with a powerful tool for disinformation[255], and as such, pose a threat to the functioning of politics and democratic societies, especially in times of societal unrest. In our example, the attackers make use of the existing anger and discontent among citizens during the Covid19 pandemic. With many people currently being unhappy about the progress of vaccinations, a corresponding deepfake that hits this notch would find fertile ground. This could lead to a lasting loss of confidence in politics.

The interesting thing about the societal impact of false political statements, however, is that even proven fake videos can be harmful. Often, when a video is taken offline or debunked after a while, the actual harm is already done. Suppose that, in our example, a verifiable video recording existed of the moment when the ministers made the alleged statements, so that they could prove that the footage was fake. Even then, many people would continue to believe the message of the deepfake video, or at least distrust the politicians, due to cognitive and psychological mechanisms. This is referred to as the 'misinformation effect' similar to the idea that 'where there's smoke, there's fire'.[256] In addition, people believe what fits their world view, even when there is proof of fakery, a phenomenon which is associated with cognitive biases as well as 'echo chambers' and 'filter bubbles'.[257]

Vaccari and Chadwick (2020) conclude that deepfakes have the potential to contribute to a 'generalized indeterminacy and cynicism, further intensifying recent challenges to online civic culture in democratic societies'. In their study, they find that people are more likely to feel uncertain about *what* to believe, rather than to be actually misled by deepfakes. But even if deepfakes may not necessarily deceive individuals, several interviewees indicate they can still provoke a large degree of uncertainty, which could reduce their trust in any news they find online, and this could cause confusion and apathy. So regardless of the immediate consequences, the medium- to long-term consequence of such a deepfake attack could be the erosion of citizens' trust in politics.

[255] One might think that deepfakes are just another mode of disinformation. Some scholars argue that deepfakes are substantially more powerful than other tools for spreading fake news, because of the visceral effect of video footage. In an online experiment (N=278), Dobber et al. (2021) found evidence that deepfakes might indeed be a more powerful mode of disinformation in comparison with the fake news stories and twitter trolls. The authors conclude that a surprising low number of participants recognized the deepfake as being manipulated, and that public awareness of deepfakes should therefore improve.

[256] Wayne Weiten, *Psychology: Themes and Variations: Themes and Variations* Wadsworth/Cengage Learning, 2010.

[257] Interview Schneider

Regulatory context

Because the creator is a capable organisation such as a foreign intelligence service operating outside of national or EU regulations, various existing regulations would be irrelevant. Furthermore, since the foreign intelligence service possesses sophisticated attack methods, both existing and future security methods could very likely be circumvented. These include, for example, the voluntary labelling of deepfake videos operated by current software manufacturers, but also deepfake or botnet detection methods of social media platforms.

Just like the scenario of deepfake pornography, it is clear that many different actors play a role in the lifecycle of this particular deepfake. They may all have different rights and responsibilities. Legal review of the case would likely find that the deepfake video interfered with the victims' personal data and privacy rights. Additionally, in most Member States, the content of the disseminated deepfake alleging false facts is likely to constitute a criminal offence of defamation. This means that they could take legal steps against the deepfake. However, since the perpetrator in the case discussed here is a foreign intelligence service that has used botnets, the question of holding the perpetrator responsible is a difficult one. The first difficulty is the identification of the creators.

Depending on what software was used to produce the deepfake, what hacking software was used to break into the victims' communications networks, and what software was used to create and operate the botnet, and if they may have been actively involved in planning or executing the attack, the software producers could also be held responsible. However, whether reliable statements can be made about such software would have to be addressed in each individual case. Presumably, the intelligence services or those of allied countries would be able to assist in the reconnaissance. If the accomplices are identifiable and known to be in another state, a request for arrest or provisional arrest with a view to extradition could also be considered. However, if the intelligence service of a foreign state has commissioned the attack, as in the present example, this measure will also come to nothing, since the perpetrators of cross-border cyberattacks usually cannot be found.

This underscores how difficult it can be to counter a sophisticated deepfake attack. Existing regulations would be insufficient to counter the negative impact of the scenario discussed here. In view of the desolate state of cyber security throughout the EU on the one hand, and the readiness of some states to attack on the other, the scenario presented here should certainly attract attention.

7.3. Manipulated court evidence

Setting the stage

Imagine you are a popular politician, swiftly climbing the country's political ladder. This upsets those who are in power, as it seems you will gain a huge victory in the upcoming elections. Then, to your shock and disbelief, an audio tape emerges. The public can hear your voice in a telephone conversation speaking to an unknown person, discussing the possibility of taking bribes. Shortly afterwards, the police arrest you. Election day passes by while you are in prison in anticipation of a criminal trial for corruption. The telephone conversation is the result of an AI-generated synthetic deepfake, devised by your political opponents. It is your voice, but you never said those things. How will you convince the judge that the audio tape is a forgery?

Unfortunately, the scenario in which a tampered audio recording leads to the labelling of a person as 'criminal', is not fictional.[258] Other forged pieces of evidence, such as falsified documents or imagery,

[258] See, for instance, the judgment of the European Court of Human Rights (ECtHR) 10 October 2019, 8284/07 (*Batiashvili v Georgia*) in which a telephone conversation was tampered with by the authorities and distributed to a television channel for broadcasting purposes, in order to present a politician as guilty. The ECtHR hold that this violated the presumption of innocence principle of article 6(2) European Convention on Human Rights (ECHR).

predate the digital era and have been used to mislead society, including the judiciary.[259] This happens in criminal proceedings, with the risk of miscarriages of justice, but also in civil proceedings. For instance, as part of a child custody case, the mother of the child tried to convince the judge that her husband behaved violently. She manipulated an audio recording of the man to make it look like he was making threats.[260] Although this was a so-called cheap fake, a less sophisticated variant of deepfakes, it raises questions and concerns. What if the manipulated recording remained unproven as fake?

Europol shares these concerns when it refers to the ongoing improvements of deepfakes in its 2019 Internet Organised Crime Threat Assessment. It notes that this technology is improving rapidly and is becoming more accessible and easier to use. The EU's law enforcement agency warns: 'This can have serious implications for law enforcement authorities, as it might raise questions about the authenticity of evidence and complicate investigations.'[261]

This short case study investigates the most prominent questions when evidence manipulated by deepfake technology enters the courtroom, either knowingly by malicious actors or even unknowingly by a party who wants to substantiate its case. What is the potential impact of this scenario? How will litigators and judges determine whether evidence is real or fake? How does the law safeguard the authenticity of evidence?

Actors involved

Before we assess the potential impact and legal context of manipulated evidence, it is important to address the relevant actors whose rights might be affected. In the first place, these are the individual whose voice has been used to create the deepfake audio fragment (**the victim**), and the actor who that created the deepfake (**the perpetrator**). The general public (**the audience**) might be misled by the audio fragment, in order to vote for other politicians or parties. Like in the other scenarios, the deepfake audio fragment has probably been created using software or a service provided by a technology developer (**technology provider**), and disseminated by individuals (**platform users**) through an online intermediary service (**the platform**). In this scenario, there was no original video or audio material used for the creation of the deepfake, meaning that there are no copyright holders involved.

Civil proceedings are about rights and liabilities. Usually there is a legal dispute between a plaintiff who has a legal claim, aimed at the other party which contests that claim. Normally, the plaintiff claims that the defendant must do something (e.g. pay in accordance with the contract) or refrain from doing something (e.g. the landlord must not evict the plaintiff). Civil proceedings can apply to all kinds of legal claims, such as the child custody case mentioned in the previous section. The court issues its judgment, based on the evidence put forward by the plaintiff and the defendant. This evidence, such as audio-visual materials, can be deepfakes, fabricated by either one of the litigating parties in order to win.

Potential impact

Deepfakes, like any material that does not depict reality, pose several risks to the judicial system. First of all, such material can have an impact on the suspect or litigating parties in civil proceedings. In the worst-case scenario, a suspect receives a sentence and may even end up in prison, based on manipulated evidence. In civil proceedings, a party may unjustifiably lose a case because its opponent used convincing evidence which was altered by deepfake technology. For instance, the judge may hold

[259] For instance, the forgery of documents has been illegal in The Netherlands at least since the enactment of the Dutch Penal Code in 1886. For an overview of forged material throughout modern history, see: Fausto Galvan and Arma Carabinieri, 'Image/Video Forensics' CEPOL, 2020, European Law Enforcement Research Bulletin.

[260] Matt Reynolds, 'Courts and Lawyers Struggle with Growing Prevalence of Deepfakes,' ABA Journal, 2020.

[261] 'Internet Organised Crime Threat Assessment (IOCTA) 2019' Europol, 2019.

that the father is aggressive based on the fake audio conversation and award sole custody to the mother. These are far-reaching consequences.

Manipulated evidence can be hard to detect. This requires technological and procedural measures to authenticate the material. Such measures are not new, as most criminal law systems mandate a so-called 'chain of custody' which chronologically 'records the sequence of custody, control, transfer, analysis, and the disposition of evidence'. However, the rules describing the measures may have to be updated in certain countries to include the detection of deepfakes and to ascertain the originality of the evidence. For instance, by maintaining procedures that counter the possibility that footage from police body cams can be altered remotely.[262]

In addition to this, legal scholars worry about the 'liar's dividend', regarding the possibility for a suspect to claim that any evidence is fake or constructed. This may hinder effective prosecution of a case. Instead of proving that the defendant committed the crime, the prosecution may now need to prove, beyond reasonable doubt, that the evidence such as an audio- or video-recording is authentic and not manipulated.[263] This can be difficult or even impossible. The scholars raise their concerns in relation to the American legal system, but their notions could also apply to similar systems in Europe.

There is also a risk of deepfakes introduced as evidence where these materials are not recognised to be inauthentic. This could raise the bar for evidence.[264] However, in civil proceedings a higher bar for authenticating audio-visual materials could lead to costly procedures, as the plaintiff or the defendant have to invest in forensic experts who can determine the authenticity of the evidence, or verify that it has not been manipulated.

Lastly, there is a wider impact that goes beyond individual court cases. Gartner predicts that by the year 2022 the majority of people will consume more false than true information.[265] As it becomes harder to tell what is real, courts and jurors may start questioning the authenticity of every piece of evidence.[266] Apart from deepfakes damning the innocent and exonerating the guilty, people may doubt unaltered content because they know realistic deepfakes are possible.[267] This has a potential corrosive effect on the justice system.[268] If deepfakes become widespread in the court room, this threatens the rule of law as the technology erodes people's trust in information.

Regulatory context

The Treaty on the European Union states that the EU shall offer its citizens an area of freedom, security and justice.[269] As part of this mission, the EU must protect fundamental rights. Relevant rights can be found in chapter V ('Justice') of the Charter of the Fundamental Rights of the EU. The ECHR is relevant as well, specifically Article 6 ('right to a fair trial').

[262] Lily Hay Newman, 'Police Bodycams Can Be Hacked to Doctor Footage,' *Wired*, 2018.

[263] Agnes Venema and Zeno Geradts, 'Digital Forensics, Deepfakes, and the Legal Process,' *The SciTech Lawyer*, 2020.

[264] Philip Boucher, 'Artificial Intelligence: How Does It Work, Why Does It Matter, and What Can We Do about It?' Panel for the Future of Science and Technology, 2020.

[265] Judit Bayer et al., 'Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States - Think Tank' Directorate General for Internal Policies of the Union, 2019.

[266] Riana Pfefferkorn, 'Too Good to Be True? 'Deepfakes' Pose a New Challenge for Trial Courts,' *Washington State Bar Association*, 2019.

[267] Agnieszka McPeak, 'The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood,' SSRN Scholarly Paper Rochester, NY: Social Science Research Network, February 21, 2021.

[268] Pfefferkorn, 'Too Good to Be True?'

[269] Article 3(2) Treaty on European Union.

Article 6 ECHR is concerned with, among other things, whether an applicant was afforded ample opportunities to contest the evidence that they considered to be false.[270] For instance, information which is essential for the assessment of the lawfulness of a detention should be made available in an appropriate manner to a suspect's lawyer.[271] However, Article 6 does not prescribe rules of evidence, such as rules on the admissibility and probative value of evidence. Member States should deal with these matters. However, in practice, their national laws regarding evidence are influenced by EU regulations and directives.

For instance, EU Directive 2012/13/EU provides a right of access to material evidence to which arrested and detained persons, or their lawyers, should have access.[272] This enables them to challenge the merits of the accusation. Evidence altered by deepfakes would be covered by the definition of material evidence, as it includes materials such as documents, photographs, audio- and video-recordings. The EU eIDAS Regulation[273] also establishes various types of digital evidence. Think of electronic seals and time stamps as well as electronic signatures. As a principle rule, these electronic means should not be denied legal effect and admissibility as evidence in legal proceedings.[274] The eIDAS Cooperation Network could play an important role in addressing the authentication and verification issues with regard to materials altered by deepfake technology.[275] Furthermore, the proposed artificial intelligence act[276] qualifies AI systems used by law enforcement to detect deepfakes as 'high-risk' AI systems.[277] Indeed, the Act recognises that the 'exercise of important procedural fundamental rights, such as the right to a fair trial as well as the right of defence and the presumption of innocence, could be hampered, where such AI systems are not sufficiently transparent, explainable and documented'.[278]

Lastly, on an international level, we refer to the Guidelines to Facilitate the use of Electronic Evidence in Court Proceedings, as adopted by the Council of Europe's Committee of Ministers.[279] This is the first such international instrument which deals with, among other things, the use of electronic evidence. However, legal scholars have already commented that some sections require elaboration regarding how the authenticity or integrity of electronic evidence can be challenged.[280] Their remarks are relevant considering deepfake material in the courtroom.

[270] Judgment of the ECtHR 6 June 2002, 53254/99 (*Karalevičius v Lithuania*).

[271] Judgment of the ECtHR 13 February 2001, 23541/94 (*Garcia Alva v Germany*).

[272] Directive 2012/13/EU of the European Parliament and of the Council of 22 May 2012.

[273] Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. See also: Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD).

[274] See: articles 25, 35, 41, 43 and 46 of the EU eIDAS regulation.

[275] European Commission, 'Communication on Tackling Online Disinformation: A European Approach,' 2018.

[276] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 21 April 2021, COM(2021) 206 final, 2021/0106 (COD).

[277] Point 6(c) of Annex III in conjunction with article 6(2) of the proposed Artificial Intelligence Act. See also points 7(c) and 8 of Annex III regarding AI systems used by competent public authorities for the verification of the authenticity of documents and to detect non-authentic documents as well as AI systems intended to assist a judicial authority in researching and interpreting facts and the law.

[278] Recital no. 38 of the proposed Artificial Intelligence Act.

[279] Guidelines of the Committee of Ministers of the Council of Europe on electronic evidence in civil and administrative proceedings, CM(2018)169-add1final, 30 January 2019.

[280] Remigijus Jokubauskas and Marek Świerczyński, 'Is Revision of the Council of Europe Guidelines on Electronic Evidence Already Needed?,' *Utrecht Law Review* 16, no. 1 (May 26, 2020): 13–20.

7.4. Conclusion

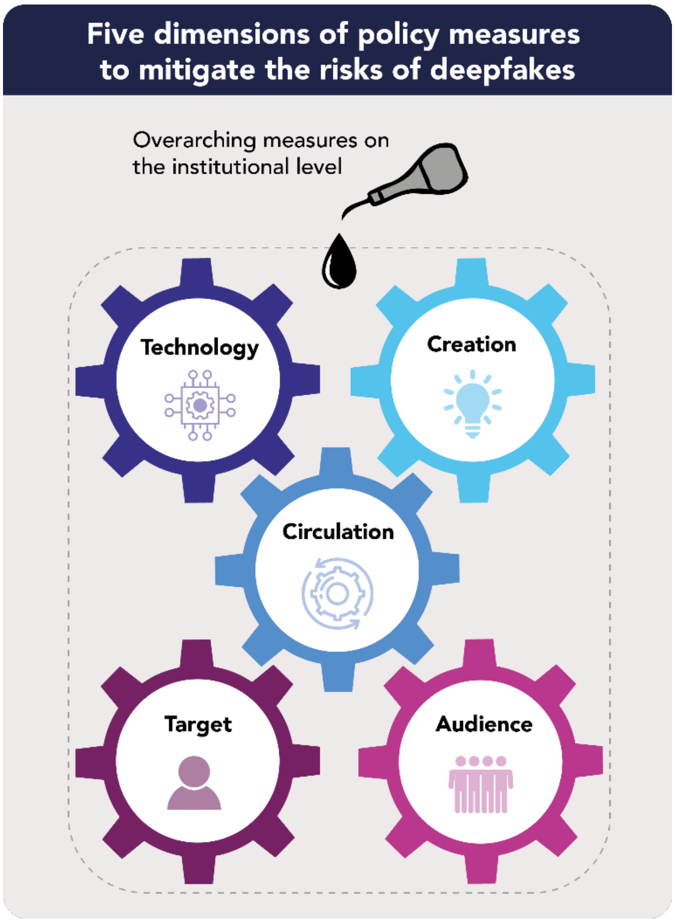
There are some important lessons to be learned from the three scenarios presented in this chapter. The scenarios illustrated how the impact of a single deepfake often exceeds the personal level. The broad societal impact of deepfakes is almost never limited to a single type of risk, but rather to a combination of cascading impacts at different levels. We have also seen that, even though the current rules and regulations offer at least some guidance for mitigating potential negative impacts of deepfakes, the legal route for victims remains challenging. Typically, there are different actors involved in the lifecycle of a deepfake. These actors might have competing rights and obligations. The scenarios illustrated how perpetrators often act anonymously, making it harder to hold them accountable. It seems that platforms could play a pivotal role in helping the victim to identify the perpetrator. Moreover, technology providers also have responsibilities in safeguarding positive and legal use of their technologies. This leads to the conclusion that when aiming to mitigate the potential negative impacts of deepfakes, policy-makers should take different dimensions of the deepfake lifecycle into account. These dimensions will be introduced in the next chapter.

8. Regulatory options

In the previous chapter, we saw that it can be challenging for victims of deepfakes to assert their rights. While current rules and regulations already set requirements and pose limits to deepfake creation and dissemination, they currently fail to prevent the impacts of malicious deepfakes, and accountability gaps remain. Moreover, we have learned that even when a video is proven to be fake, and taken offline after a while, the individual and societal harm is often already done. This implies that tackling the negative impacts associated with deepfakes requires both a preventative and a reactive approach, aimed at avoiding and addressing undesirable uses of deepfake technology.

In this chapter, we identify various policy options for mitigating the negative impacts associated with deepfakes. In line with the different phases of the 'deepfake lifecycle', we distinguish five dimensions of policy measures. These include: 1. Technology, 2. Creation, 3. Circulation, 4. Target, 5. Audience. In addition to these five dimensions, we also identify some overarching measures that could be considered on the institutional and organisational level. We present all the options that emerged from our analysis based on the literature study, expert interviews and reviews, without favouring one over the other. Whenever we identify downsides to a particular policy option, we mention them below.

Figure 10 - Five dimensions of policy measures to mitigate the risks of deepfakes


Five dimensions of policy measures to mitigate the risks of deepfakes
Overarching measures on the institutional level
Technology
Creation
Circulation
Target
Audience

Technology dimension

This dimension covers policy options aimed at addressing the technology underlying deepfakes – AI-based machine learning techniques – and the actors involved in producing and providing this technology. The regulation of such technology is largely the domain of the AI regulatory framework as proposed by the European Commission. The framework takes a risk-based approach to the regulation of AI. As described in Chapter 6, deepfakes are explicitly covered in the Commission proposal as 'AI systems used to generate or manipulate image, audio or video content', and have to adhere to certain minimum requirements, most notably when it comes to labelling. They are not included in the 'high risk' category, and uncertainty remains whether they could fall under the 'prohibited' category. The current AI framework proposal thus leaves room for interpretation. This report has documented a wide range of applications of deepfake technology, some of which are clearly high-risk. As such, policy-makers might consider some clarifications and additions to the AI framework proposal.

Policy options within the AI framework proposal:

- **Clarify which AI practices should be prohibited under the AI framework:** The proposed AI framework currently mentions four types of prohibited AI practices that could relate to certain applications of deepfake technology. However, the formulation of these sections are unclear and open to interpretation. Based on the findings in this report, we conclude that some deepfake applications would fulfil the conditions mentioned in article 5(1)(a) and (b). For example, deepfakes that enable a deceptive manipulation of reality, or are capable of inciting violence against people or causing violent social unrest. However, since deepfakes are explicitly mentioned in article 52(3), it is unclear whether they could also be covered under article 5(1) lit a and b.[281] The clarification of this matter by the European Commission may be appropriate.
- **Create legal obligations for deepfake technology providers:** In the present AI framework proposal, there are no obligations for technology providers to enable the labelling of deepfake content. Article 52(3) does however pose obligations on the user. This means that in the current AI proposal, the responsibility for labelling deepfakes lies fully with the creator of a deepfake video. A policy option could be to extend the AI framework proposal by obliging technology producers (providers) to incorporate labelling features.
- **Regulate deepfake technology as high-risk:** Another option in this respect could be to regulate deepfake technology as a high-risk AI system under the AI framework proposal, and thus add deepfake technology to annex III. Looking at the long list of risks and adverse impacts associated with deepfakes in Chapter 5 of this report, the case can easily be made that the AI technology underlying deepfakes can impact fundamental rights and safety; the criterion used by the European Commission to determine AI systems as high risk. If deepfake technology was to be treated as high risk, explicit legal requirements would be placed on the provider of the technology, including carrying out risk-assessments, providing for documentation and human oversight, and ensuring high-quality datasets.
- **Place limits on the spread of deepfake detection technology:** Detection technology is crucial in stopping the circulation of malicious deepfakes. However, if deepfake technology providers are aware of the detection technologies, they can adjust the deepfake production technologies and circumvent detection. This leads to a cat-and-mouse-game between deepfake production technology and deepfake detection technology. One way to address this is to place greater limits on the spread of cutting-edge deepfake-detection technology/research by technology providers, so that advances made by digital forensics researchers are not immediately neutralised by their adversaries.[282] This policy option is a controversial measure, because by limiting the use of detection technology to, for example, law enforcement agencies, other actors may remain unable to detect deepfakes and suffer the consequences described in Chapter 5. The up- and downsides of this measure should thus be carefully weighed before being adopted.

[281] European Commission, 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).'

[282] Collins, 'Forged Authenticity' p. 17.

Policy options beyond the AI framework proposal:

- **Invest in the development of AI systems that prevent, slow or complicate deepfake attacks:** Even though a technological response to malicious use of deepfake technology will not be sufficient to address all the risks and impacts associated with the technology, there is a role for AI to detect and prevent deepfake attacks (see Section 3.6 for potential and limits of these types of technologies). Investments in the development of AI systems that prevent, slow, or complicate deepfake attacks are therefore advisable. The European Commission could consider including this as a focus area under Horizon Europe.
- **Invest in education and raise awareness of AI professionals:** The adverse impacts associated with AI, including deepfakes, could become a standard part of the curriculum of AI researchers and AI developers, and prove to be a valuable way to teach IT professionals about the ethical and societal effects of their systems.[283] AI ethics and impact assessments could become part of educational policies at European, Member State and institutional level

Creation dimension

This dimension covers the policy options aimed at addressing the creator of deepfakes, or in AI framework terminology: the 'users' of AI systems. The AI framework proposal already formulates some rules and restrictions for the use of deepfake technology, but additional measures are possible.

This dimension also addresses those who use deepfake technology for malicious purposes: the 'perpetrator' described in Chapter 7. As we have seen in previous chapters, malicious users of deepfake technology often hide behind anonymity and cannot be easily identified, thereby escaping accountability. These users cannot be expected to willingly comply with the labelling requirement as introduced in the AI framework proposal. Policy measures needed against malicious users of deepfake technology therefore go beyond this labelling requirement.

Policy options within the AI framework:

- **Clarify the guidelines for the manner of labelling:** In the proposed AI framework it is unclear what labelling exactly entails. What information needs to be provided, and how should it be presented? Is there a standard, or is it up to the user to decide how to label? More guidance on the manner of labelling could be added to the AI framework proposal. From the perspective of the audience (the receiver of the deepfake), standardised labelling could be preferable, since receivers will then know what to look for.
- **Limit the exceptions for the deepfake labelling requirement:** The AI framework proposal places a labelling obligation on users of deepfake technology, but creates exemptions when deepfakes are used for law enforcement, as well as in arts, sciences and where the use 'is needed for freedom of expression'.[284] These exceptions are so broad and open to interpretation that, as a result, many deepfakes may remain unlabelled, and a discussion on the labelling requirement before the courts can be predicted. Weighing the potential negative impacts of deepfakes when they are not recognised as such against their beneficial use in arts, sciences and expression, one could argue that labelling in these situations is recommendable.[285] Would a deepfake-based

[283] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario'; Maria Pawelec, Cora Bieß, and Alexander Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern,' Eberhard Karls Universität Tübingen, 2021.

[284] Article 52 (3) European Commission, 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).'

[285] Pawelec, Bieß, and Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern.'

artwork or a satirical deepfake video protected by the freedom of expression be less valuable when the deepfake character is revealed?

- **Ban certain applications:** Considering the potential negative impact of specific applications of deepfakes (such as non-consensual deepfake pornography or political disinformation campaigns), transparency obligations alone seem insufficient to address such negative impacts. A full ban of deepfake technology on the other hand also seems disproportionate, considering the dual use character of deepfake technology. A policy option could be to prohibit certain types of applications. In some countries, such proposals have already been put forward, for example in the United States of America (USA – see Section 6.8.1), the Netherlands and the UK.[286] This policy option should be weighed carefully, since banning certain applications will always be accompanied by a deterrent effect with regard to freedom of expression.
- **Ban deepfake political advertising or communication:** Research indicates that deepfakes combined with micro-targeting may be used to manipulate political opinions. In order to mitigate this risk, policy-makers could consider extending the European democracy action plan (EDAP) that was announced by the European Commission in December 2020.[287] The European Commission already aims to introduce measures to increase the transparency of political advertising and communication, including measures against micro-targeting in general. Given the possible strong manipulative effect of deepfakes, policy-makers could consider including a complete ban on the use of deepfake technologies in such advertisements or communication.

Policy options beyond the AI framework proposal:

- **Extending the current legal framework with regard to criminal offences:** Considering the harm that malicious uses of deepfakes may cause, an assessment of the robustness of existing rules and regulations at Member State level could be helpful to assess whether the addition and specification of existing criminal offences is necessary/desirable. In Germany, for example, the distribution of a deepfake that violates personal rights (such as deepfake pornography) is prohibited, but not its production.[288] In addition, the introduction of a criminal offence of impersonating a person with intent to 'harm, intimidate, threaten, or defraud', as is the case in some US states, could also be useful (see Section 6.8.1). The basis for this could be an EU-wide comparative study of the national legal situation, which could be promoted within the framework of Horizon Europe. Based on this, the Member States could then take action.
- **Diplomatic actions and international agreements to refrain from the use of deepfakes** by foreign states and their intelligence agencies. The use of disinformation and deepfakes by foreign states, or actors associated with foreign state institutions, contributes to increasing geopolitical tensions. In order to prevent and de-escalate conflicts, there is a need for intensified diplomatic actions and international cooperation.[289] Although some successes have been achieved regarding agreements at a regional level,[290] binding global agreements that deal with information conflicts and the spreading of disinformation have so far not been made.
- **Impose economic sanctions on states engaged in disinformation and deepfakes:** As mentioned above, states may actively use deepfakes in disinformation campaigns. Sometimes the creation of a deepfake might be traced back to a specific perpetrator that can be linked to a

[286] 'Petition: Criminalise Manufacturing and Distributing Deep-Fake Pornography,' Petitions - UK Government and Parliament, 2021; 'Politiek en meldpunt binden strijd aan met 'deep nudes,'' NOS, 2020; 'Rathenau Manifesto: Set 10 design requirements for tomorrow's digital society now,' Rathenau Instituut, 2020.

[287] COM(2020)0790 https://ec.europa.eu/info/sites/default/files/edap_communication.pdf

[288] Tobias Lantwin, 'Deep Fakes – Düstere Zeiten Für Den Persönlichkeitsschutz? Rechtliche Herausforderungen Und Lösungsansätze,' *MultiMedia Und Recht*, 2019.

[289] Jurriën Hamer, Rinie van Est, and Lambèr Royakkers, 'Cyberspace without Conflict' Rathenau Instituuut, 2019.

[290] Such as the IMPACT coalition, the European network of Cyber Emergency Incident Response Teams and the NATO cyber exercises (Hamer et al, 2019)

foreign state institution. If diplomacy does not yield sufficient results, a policy option is to impose well-considered economic sanctions.[291]

- **Critical discussion of the measure to lift anonymity for using online platforms:** In China, users of online platforms need to register with their identity (ID) before being able to enter platforms. The discussion as to what level of anonymity is acceptable and desirable online is a sensitive one. On the one hand, platform user anonymity provides cover for malicious users. On the other hand, anonymity serves as a useful protection for activists and whistleblowers. Possible approaches exist that only require user identification before uploading certain content. Weighing these different options including their up- and downsides deserves careful attention. Both the European Commission and Member States could encourage further public debate, as well as promote studies on possible consequences.
- **Invest in knowledge and technology transfer to developing countries:** As the possible negative impact of deepfakes in developing countries is regarded as greater than in developed countries,[292] knowledge and technology transfer to these countries can help improve the resilience of developing countries against the risks of deepfakes. The European Commission and Member States could consider incorporating such knowledge and technology transfer into foreign and development policies.

Circulation dimension

This domain covers the policy options aimed at addressing the circulation of deepfakes, by formulating possible rules and restrictions for the dissemination of (certain) deepfakes. As the case scenarios in Chapter 7 have demonstrated, online platforms, media and communication services play a crucial role in the dissemination of deepfakes. The dissemination and circulation of a deepfake determines to a large extent the scale and the severity of the impact. Therefore, responsibilities and obligations for platforms and other intermediaries are often recommended, including the liability of platforms if they fail to fulfil their obligations. The proposed digital services act, in development at the time of writing, provides a window of opportunity to take measures to limit the circulation of deepfakes.

Policy options for the digital services act:

- **Detecting deepfakes:** Because of the central role online platforms and other intermediaries play in the dissemination of deepfakes, policy-makers could consider obliging platforms and other intermediaries to have deepfake detection software in place as a prerequisite for possible labelling. An alternative for detection is the **use of upload filters** (e.g. image rights and control: faces are blurred, until consent is given by the persons depicted).[293] The downside of these upload filters is that they can become too restrictive and limit freedom of expression.
- **Detecting authenticity:** Platforms could also be obliged to deploy methods to detect the authenticity of online identities[294] and to detect bot(-net)s.[295] This would enable identification of fake accounts and artificial amplification, which have been proven to play a significant role in disseminating deepfakes and disinformation.
- **Establish labelling and take-down procedures:** Policy-makers could consider obliging platforms to label detected deepfakes as such and/or to take down unlabelled deepfakes once the platform is notified by a victim or trusted flaggers following established procedures. In order to guarantee the fairness of take down decisions, these decisions should be accompanied by human oversight (no algorithmic automatic take-down), be transparent, and accompanied by

[291] Chesney and Citron, 'Deep Fakes,' 52.

[292] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario,' 27.

[293] Collins, 'Forged Authenticity,' 18.

[294] Interview Thies

[295] Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario,' 31.

informing the user about the decision, to allow an opportunity for appeal. Furthermore, measures would need to be taken to limit abuse of these procedures, for example, by only authorising rights holders or entities acting on their behalf to report copyright infringements. Furthermore, in the context of deepfakes, a distinction could be made between reporting by any person and reporting by the affected persons themselves. A report in the latter case could then lead to prioritised processing of the concern.[296]

- **Oblige platforms to have an appeal procedure in place:** To allow for rectification of unjustified labelling or take down by platforms and intermediaries, platforms could be obliged to have an accessible and transparent appeal procedure in place.
- **Limit the decision-making authority of platforms to decide unilaterally on the legality and harmfulness of content:** From a human rights perspective, it is problematic to leave the decision on what content is illegal or harmful unilaterally in the hands of the platforms. Independent oversight of content moderation decisions seem important to limiting the influence of online platforms and intermediaries on freedom of expression and the quality of social communication and dialogue.[297] The further development of the DSA offers European institutions an opportunity to provide guidance in designing these oversight mechanisms.
- **Increase transparency:** To monitor the occurrence of deepfake dissemination, as well as the effectiveness and fairness of detection policies, policy-makers can consider adding to the transparency obligations in the DSA reporting obligations on their deepfake detection systems, detection results and decisions.
- **Slow down the speed of circulation:** While freedom of speech is a fundamental right, freedom of reach is not. That is to say, our freedom of expression does not include an automatic entitlement to widespread distribution of what we say[298] Policy-makers could consider obliging platforms to take measures to slow down the speed of circulation, for instance by:
    - Limiting the number of users in (chat) groups[299]
    - Reducing the speed and the dynamics at which content can be shared or platform nudges,[300] e.g. introduction of a reflection period[301]
    - Restricting the possibilities for micro-targeting to reduce the risk of addressing deepfakes directly to a susceptible audience.

  These obligations could be placed under the digital services act or within the context of the European democracy action plan.

Target dimension

As we have seen in Chapter 5, malicious deepfakes create impacts at the individual level, for the person(s) depicted in the deepfake. Digitised attacks might have different and stronger effects than traditional patterns of crime, e.g. an abuse of an image online can cause a much longer lasting harm on individuals than the same crime in the offline world because, for example, it is seen by many more people.[302] The case scenarios have demonstrated that the rights of victims may be protected on paper, but it often proves hard to enact these rights. We therefore offer several options to improve the protection of the victim.

[296] Julia Reda, 'Der digital services act steht für einen Sinneswandel in Brüssel,' *Netzpolitik.org* (blog), 2021.

[297] Pawelec, Bieß, and Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern'; 'De Toekomst van Online Platforms' Rathenau Instituut, 2021.

[298] Renee Diresta, 'Free Speech Is Not the Same As Free Reach,' *Wired*, 2018.

[299] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy,' 88.

[300] Collins, 'Forged Authenticity'; Pawelec, Bieß, and Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern.'

[301] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy.'

[302] Collins, 'Forged Authenticity,' 21.

Policy options:

- **Institutionalise support for victims of deepfakes:** Victims of deepfakes can experience difficulties in finding out what avenues for remedy and justice are available, because of the complexity of applicable rules and regulations and the multiple actors involved, combined with the vulnerability of the victims in the face of often anonymous perpetrators. Policy-makers could consider setting up or tasking an existing advisory body for easily accessible judicial and psychological support at Member State level.[303] This body could help victims navigate avenues for justice (e.g. notify platforms for take-down of deepfakes, assist in identifying the perpetrator, start civil action against perpetrators, incite criminal action against perpetrators), as well as avenues for psychological support. This advisory body could also play a role in monitoring the occurrence of deepfakes and informing policy options to better protect victims in the future.
- **Strengthen capacity of data protection authorities to respond to the use of personal data for deepfakes:** As we have seen in Chapter 6, the creation of deepfakes almost always involves the processing of personal data, both for training the algorithms in the technology and for creating a specific deepfake. It is up to the data protection authorities to respond to the reports of unlawful data processing, and the deepfake phenomenon introduces a whole new category of such unlawful processing. Data protection authorities should be equipped with resources to respond to the challenges posed by deepfakes. Because the next evaluation of the GDPR is far in the future, the European Data Protection Board (EDPB) could be called upon to review the implementation of the GDPR requirements to determine whether and to what extent the existing resources of the Member State data protection authorities are sufficient to address the challenges posed by deepfakes.
- **Provide guidelines on the application of the GDPR framework to deepfakes:** The EDPB could develop guidelines on how the GDPR framework applies to deepfakes, by clarifying what is and is not permitted under the current GDPR framework. These guidelines could clarify for example under which circumstances a data protection impact assessment is required, and how personal data protection relates to freedom of expression in the context of deepfakes.
- **Extend the list of special categories of personal data with voice and facial data:** Deepfake technology processes personal data, since it uses personal features such as voice or facial landmarks to identify the targeted individual. In the next revision of the GDPR, the European Commission could extend the list of special categories of personal data with voice and facial data, to respond to deepfakes as an upcoming phenomenon. This extension would further clarify the balance between protection of personal data and freedom of expression when it comes to deepfakes, as well as clarify the exceptions under which the use of voice and facial data is permitted.
- **Develop a unified approach for the proper use of personality rights within the European Union:** Personality rights are recognised by many different laws, including various rights of publicity, privacy and dignity, with complicated exceptions. This may lead to unpredictable outcomes in lawsuits and different outcomes per Member State. The 'right to the protection of one's image' could be further developed and clarified in this regard on EU level.
- **Protect personal data of deceased persons:** Deepfakes can be used to virtually revive deceased persons and make it appear as if they said or did things they did not, without their consent. This raises ethical and fundamental rights questions. Similar to a donor codicil, in which people register whether they want to donate their organs after their death. The introduction of a data codicil, in which people declare how they want their data to be used after their death, could be considered. Furthermore, the GDPR and/or its national equivalents could be expanded by including protection of personal data of deceased persons, in line with current measures in

[303] Pawelec, Bieß, and Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern.'

Hungary and Spain (see paragraph 6.8.2). This could be included as a topic for the next round of evaluation of the GDPR.

- **Address authentication and verification procedures for court evidence:** The EU eIDAS Regulation[304] establishes various types of digital evidence (e.g. electronic seals and time stamps and electronic signatures). As a principle rule, these electronic means 'should not be denied legal effect and admissibility as evidence in legal proceedings'.[305] The eIDAS Cooperation Network could play an important role in addressing the authentication and verification issues with regard to materials altered by deepfake technology.[306] For example, by providing explanatory guidance to courts on how to deal with arguments that challenge the authenticity of digital evidence.

Audience dimension

As we have seen in Chapter 5, impacts transcend the individual level and can cascade to group or even societal levels. Whether that will happen, partly depends on the audience response: will they believe the deepfake, disseminate deepfakes further when they receive them, lose trust in institutions? Therefore, we have termed the final crucial dimension for policy-makers to limit the risks and impacts of deepfakes as the audience dimension.

Policy options:

- **Establish authentication systems:** As an alternative to the labelling of deepfakes, consider establishing systems that help the receiver of a message to verify the authenticity of the message. Options mentioned by experts are the use of raw video data (no hidden fakes), the use of technologies to prove authenticity of videos (digital watermarks) or registering the provenance of information to allow for traceability of sources.[307] However, non-intended consequences of authentication systems should also be considered, most importantly when it risks adversely impacting the safety of journalists and whistleblowers.[308] Authentication systems could also create unnecessary barriers for (citizen) journalism.
- **Invest in media literacy and technological citizenship:**[309] Awareness and literacy of deepfake technologies could increase the resilience of the public against the risks of deepfakes. Such efforts would ideally include the general public, as well as organisations and institutions (for example non-state actors such as companies and organisations, and state-institutions, such as supervisory bodies that might be particularly affected by deepfake attacks.[310] This means investing in education across a wide array of actors, beginning at the primary school level and continuing in professional training, for example with regard to journalists and AI professionals. Teaching quick and easy debunking strategies for malicious deepfakes is also helpful for the general public. For example, in the event that doubts arise about the authenticity of a counterpart during a telephone conversation, people can be trained to hang up and call or contact the person again (via another channel). Furthermore, it could be helpful if social media users were more (self-)critical towards the content they consume and share, for example by making it a habit to not share potentially problematic videos without checking their authenticity. Policy-makers can

[304] Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. See also: Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters, COM/2018/225 final - 2018/0108 (COD).

[305] See: articles 25, 35, 41, 43 and 46 of the EU eIDAS regulation.

[306] European Commission, 'Communication on Tackling Online Disinformation: A European Approach.'

[307] Collins, 'Forged Authenticity.'

[308] Pawelec, Bieß, and Orlowski, 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern.'

[309] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy.'

[310] van Boheemen, Munnichs, and Dujso, 'Digital Threats to Democracy,' 90; Bateman, 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario,' 31; Collins, 'Forged Authenticity,' 22.

further take inspiration from the approach taken by Taiwan in this regard (see Section 6.8.4), where the public is actively engaged in the process to fight disinformation themselves

- **Continue to invest in a pluralistic media landscape and high-quality journalism:** As recognised in the European democracy action plan,[311] a pluralistic media landscape is a prerequisite for access to truthful information, and to counter disinformation.[312] Funding and other support of European and national journalism and media pluralism by the European Commission and Member States remains crucial moving forward.

Institutional and organisational measures

In addition to the above five dimensions of which policy-makers should be aware when aiming to limit the adverse impacts associated with deepfakes, several overarching measures on the institutional and organisational level can be considered. These measures are meant to inform continuous learning, adaptation, oversight and resilience, to cope with deepfake challenges in the future.

Additional measures:

- **Systematise and institutionalise the collection of information with regard to deepfakes:** Systematic collection of data with regard to the development, detection, circulation and impact of deepfakes can inform the further development of policies and standards, enable institutional control of deepfake creation, and may create a change in the deepfake creation culture. This option fits well within the context of the European democracy action plan and the European action plan against disinformation. The European Digital Media Observatory currently being formed may be fit for this purpose.[313] In addition, ENISA, in cooperation with the EDPB can play a role in systematic collection of data about deepfakes (see Section 6.8.1). The European Commission could support research into both the production and detection/authentication of deepfakes.
- **Protecting organisations against deepfake fraud:** Deepfake technology represents a new type of corporate fraud. Manipulated video, audio or text can be used to impersonate both customers and colleagues, and could result in severe reputational or financial harm. Organisations therefore need to be prepared for and protected against the possibility of a deepfake fraud attack. To increase organisational resilience, companies could be stimulated to perform risk assessments across the board. Organisations could prepare their employees, consider an adequate response strategy, and strengthen their authentication processes. Public organisations, such as identity document issuers, should make sure that their verification processes account for potential image forgeries. Instead of accepting premade photographs from applicants, the only feasible trustworthy way of obtaining a picture portrait may be to ensure an officer takes the picture on the spot with a certified camera. Regardless of whether an organisation has been confronted with deepfakes or not, whenever a business process is dependent on the authenticity of audiographic material, an impact assessment of this new technology should be performed.
- **Ensure further research on deepfakes:** Since deepfakes are a relatively new phenomenon with the potential of high risks and impacts, further research into the development, detection, circulation and impact of deepfakes within the European Union seems recommended. The European Commission could consider increasing funding for research by media platforms, civil society organisations and academia in the area of detection, prevention and responses to deepfakes.

[311] European Commission, 'European Democracy Action Plan.'

[312] Rathenau Instituut, 'Digitalisering van het nieuws: online nieuwsgedrag en personalisatie in Nederland.,' 2018.

[313] 'EDMO – United against Disinformation,' accessed May 20, 2021.

Summary of policy options

Some of the options mentioned above based on the literature study and interviews are already covered by the proposed AI framework and the proposed digital services act; others would require further specification or expansion of the frameworks, and still others go beyond the scope of these two bodies of EU regulation. The table below provides an overview of all the options set out in the present study and indicates into which EU legislation or other level of governance these policy options could be incorporated.

Table 3 - Overview of policy options combined with the policy frameworks and/or governance levels where they could be addressed

Policy options	Cover in/consider by:
TECHNOLOGY DIMENSION	
Clarify which AI practices shall be prohibited under the AI framework	AI framework
Create legal obligations for deepfake technology providers	AI framework
Regulate deepfake technology as high-risk	AI framework
Place limits on the spread of deepfake detection technology	AI framework[314]
Invest in the development of AI systems that prevent, slow or complicate deepfake attacks	Horizon Europe
Invest in education and raise awareness of AI professionals	Educational policies at European, Member State and institutional level
CREATION DIMENSION	
Clarify the guidelines for the manner of labelling	AI framework
Limit the exceptions for the deepfake labelling requirement	AI framework
Ban certain applications	AI framework
Ban deepfake political advertising and communication	European democracy action plan
Extending current legal framework with regard to criminal offences	Horizon Europe and/or Member State level
Diplomatic actions and international agreements to refrain from the use of deepfakes	Member State and EU foreign policy level
Impose economic sanctions on states engaged in disinformation and deepfakes	Member State and EU foreign policy level

[314] This option is associated with some serious downsides, which should be carefully weighed before adoption, see explanation in text in Chapter 8.

Policy options	Cover in/consider by:
Critical discussion of the measure to lift anonymity for using online platform	Public debate and research at European and/or Member State level
Invest in knowledge and technology transfer to developing countries	Foreign and development policies at European and/or Member State level.
CIRCULATION DIMENSION	
Oblige platforms to have deepfake detection systems in place	DSA
Oblige platforms to have systems in place to detect authenticity	DSA
Establish labelling and take-down procedures	DSA
Oblige platforms to have an appeal procedure in place	DSA
Limit the decision-making authority of platforms to unilaterally decide on the legality and harmfulness of content	DSA
Increase transparency	DSA
Slow down the speed of circulation	DSA
TARGET DIMENSION	
Institutionalise support for victims of deepfakes	Member State level
Strengthen capacity of data protection authorities to respond to the use of personal data for deepfakes	EDPB to review implementation of GDPR in relation to deepfakes
Provide guidelines on the application of the GDPR framework to deepfakes	EDPB to provide guidelines
Extend the list of special categories of personal data with voice and facial data	Revision of GDPR
Develop a unified approach for the proper use of personality rights within the European Union	Horizon Europe, followed by EU harmonisation process
Protect personal data of deceased persons	Member States and evaluation of GDPR
Address authentication and verification procedures for court evidence	eIDAS Cooperation Network
AUDIENCE DIMENSION	
Establish authentication systems	Stakeholders
Invest in media literacy and technological citizenship	Multiple governance levels, including Member States and the European democracy action plan
Continue to invest in a pluralistic media landscape and high-quality journalism	Member State and EU level
INSTITUTIONAL MEASURES	

Policy options	Cover in/consider by:
Systematise and institutionalise the collection of information with regard to deepfakes	European democracy action plan and/or the European action plan against disinformation
Protecting organisations against deepfake fraud	Stakeholders
Ensure further research on deepfakes	Horizon Europe and research programmes at Member State level

9. Conclusions

Deepfakes accelerate the erosion of trust

Technical breakthrough innovations in AI, especially GANs, have led to the emergence of deepfakes: manipulated or synthetic audio and visual media that seem authentic, which feature (a) person(s) that appear(s) to say or do something they have never said or done. The barriers to access and application of deepfake technologies are lowering rapidly. Smartphone apps that require no technical know-how already enable anyone to make more or less convincing deepfakes. High-quality deepfakes – those essentially undetectable to the human eye – often still require significant technical skills and equipment, but this will likely change in the near future.

The rapid improvement of deepfake technologies has severe consequences for the trustworthiness of all audiographic material. It gives rise to a wide range of potential societal and financial harms, including manipulation of democratic processes, the economy, justice and scientific systems. Deepfakes enable all kinds of fraud, in particular those involving identity theft. Individuals – especially women – are at increased risk of defamation, intimidation and extortion, as deepfake technologies are currently predominantly used to swap the faces of victims with those of actresses in pornographic videos.

The increased likelihood of fakes forces society to adopt a higher level of distrust towards all audiographic information. Audiographic evidence will be confronted with higher scepticism and will have to meet higher standards. This will also mean that authentic materials can be more easily discredited. Therefore, deepfakes accelerate an already ongoing erosion of trust.

Deepfake technologies are dual-use

This research has identified numerous malicious as well as beneficial applications of deepfake technologies. The technologies offer opportunities to (cinematographic) artists, educators, advertisers and technology companies to create more engaging and personalised digital experiences. In the medical field, there are therapeutic applications in development and the technology may even give a voice to the mute. There are also many innocent applications, such as beauty filters in camera apps, and other entertaining applications for (live) video footage.

The use of deepfake technologies become problematic when the creator aims to deceive an audience with nefarious intent or impact. In practice, this may be difficult to anticipate, as a deepfake created for satire could easily be taken out of context, for example. Some applications, such as the fabrication of court evidence, defamation by non-consensual pornographic videos, or the creation of false political statements, are categorically high-risk. Measures that may be adequate for low-risk and benign deepfake applications, such as adding a label or requiring transparency on its provenance will not suffice. Non-consensual pornography can be damaging, even when labelled, for example. Malicious actors could also easily create materials without labels, or remove labels.

We conclude that the risks that deepfake technologies pose to society are serious, but context-specific. The technologies at hand are dual-use and should be regulated as such. Chapter 8 proposes several options to address this complexity.

Regulating the technological dimension of deepfakes will not suffice

Taking an AI-based approach to mitigating the risks posed by deepfakes will not suffice for three reasons. First, other technologies can be used to create audiographic materials that are effectively similar to deepfakes. Most notably, 3D animation techniques may create very realistic video footage. In our research, we discovered multiple instances in which 3D animation techniques are combined with AI-based deepfake technologies to create highly convincing videos.

Second, the potential harms of the technology are only partly the result of the deepfake videos or underlying technologies. Several mechanisms at play are equally essential. For example, for the manipulation of public opinion, deepfakes need not only to be produced, but also distributed. Frequently, the policies of media broadcasters and internet platform companies are instrumental to the impact of deepfakes.

Thirdly, although deepfakes can be defined in a sociological sense, it may prove much more difficult to grasp deepfake videos and their underlying technologies in legal terms. There is also an inherent subjective aspect to the seeming authenticity of deepfakes. A video that may seem convincing to one audience, may not be at all credible to another, as people often use contextual information or background knowledge to make a judgement about authenticity.

Similarly, it may be practically impossible to anticipate or assess whether a particular technology may or may not be used to create deepfakes. One has to bear in mind that the risks of deepfakes do not solely lie in the underlying technology, but largely depend on the usage of the societal practice in which the fabricated material is used.

Policy measures that address the technology underlying deepfakes are necessary to ensure that deepfake applications will be developed and used in accordance with EU values and fundamental rights. However, to mitigate the risks posed by deepfakes, policy-makers could also consider options that address the wider societal context and go beyond regulation of technology. In addition to the technological dimension, this research identified four further regulatory dimensions that should be taken into account: Creation; Circulation; Target; and Audience (see Figure 10). When considering all five dimensions of the lifecycle of deepfakes, a full regulatory landscape starts to form. The analysis of this landscape demonstrates that although many existing regulations seem to affect deepfakes, many gaps remain. Chapter 8 therefore contains several policy options that extend existing regulations to all five dimensions.

Citizens need additional support to protect their rights

Fraud, defamation, extortion and intimidation are already criminal offences, and intentional deception is already against common codes on professional integrity. The GDPR already offers guidance for tackling unlawful deepfake content, and a person's image may already be protected by intellectual property rights. There are therefore already existing procedures that could prevent or deal with the harms caused by deepfakes. At a glance, it may even seem that it is simply a matter of enforcement.

However, this research has shown that citizens need additional support to use their rights. The internet consists of a complex web of technological products and services, simultaneously involving actors in a multitude of jurisdictions. For an individual, it may often be very difficult to identify those that bear responsibility for harm, let alone hold them to account.

Visual manipulation is here to stay

Detection technology is crucial in halting the circulation of deepfakes. However, the development of deepfake technologies and forensic detection techniques is a cat-and-mouse game. The AI-technologies used to create deepfakes themselves benefit from superior detection techniques, because they are able to swiftly learn and adjust. This constant improvement cycle will continuously lead to ever more difficulty in detecting forgeries. It is therefore very likely that it will be impossible for a human being to identify a deepfake video without detection tools. And detection tools will always – by definition – only work for a limited period of time, until the production technologies re-adjust.

Policy-makers could therefore also focus on improving resilience in a changing media ecosystem that cannot always exclude corrupted information. Media literacy efforts should focus less on trying to identify deepfakes, and more on the skills that individuals and institutions need to obtain to construct a trustworthy image of reality, given the fact that they will be inevitably confronted with deceptive information.

Deepfake technology is a fast-moving target. It is impossible to predict precisely which way the technology will develop in the years to come. However, we can be sure that visual manipulation is here to stay. There are no quick fixes. Mitigating the risks of deepfakes thus requires continuous reflection and permanent learning. The European Union could play a leading role in this process.

Annex 1 – List of interviewed experts and reviewers

First Phase:

Interview partner	Date	Organisation	Position
Prof. Dr. phil. Ingrid Schneider	12-01-'21	University Hamburg	Professor
Sam Gregory	12-01-'21	WITNESS	Programme Director
Nic Newman	13-01-'21	Reuters Institute for the Study of Journalism	Senior Research Associate and lead author of the Reuters Digital News Report
Giorgio Patrini	13-01-'21	Private research company Sensity	Co-founder and Chief Scientist
Justus Thies	14-01-'21	TU Munich	Postdoctoral Researcher at TUM (Visual Computing Lab)
Jon Bateman	15-01-'21	Carnegie Endowment for International Peace	Fellow

Second Phase:

Interview partner	Date	Organisation	Position
Philipp Amann	16-03-'21	Europol	Head of Strategy European Cybercrime Centre
Kelsey Farish	12-04-'21	DAC Beachcroft	Solicitor
Angelica Fernandez	30-04`21	University Luxembourg	PhD

Reviewers:

Reviewer	Date	Organisation	Position
Prof. dr. Claes de Vreese	20-05-'21	University of Amsterdam	Professor
Laura Smillie	20-05-'21	Joint Research Centre	Policy Analyst
Dr. Bart van der Sloot	21-05-'21	Tilburg University	Senior Researcher

Annex 2 – Interview questions first phase

Part	Questions
General	
➢ Introduction	Description of position, area of responsibility, professional and educational background; what is the relation to deepfakes?
Deepfake Technology	
➢ Basic view on deepfakes	How do you define Deepfakes – what is your understanding?
	What are concrete technologies, use cases or examples/ applications in your context?
➢ Opportunities and Challenges	Can you describe the benefits of the AI application?
	What are the challenges? What are your concerns? What are possible risks, damages or harms?
	Are these risks specific for Deepfakes? Is there a new kind of threat?
	Which specific area is affected and who is especially affected?
	Could you identify use cases / applications with a high-risk potential but also low-risk potential?
	What are adequate prevention technologies?
	Which context factors contribute / increase / decrease the impact of Deepfakes? (Distribution of news via media platforms, mass and velocity aspects, digital transformation, technological innovations) How do human perception and behaviour contribute to the Deepfake effects? (Perception of images, media consumption)
➢ Future Developments	Deepfakes are at a stage mainly discussed with regard to visual fakes (photo, video). How far are also other kinds of fakes prepared and developed, e.g. auditive Deepfakes answering to phone calls (google duplex)? What are future options and trends with regard to other types of Deepfakes, e.g. avatar fakes?
	What is the potential that other AI technology might be affected by Deepfakes, e.g. autonomous mobility by fakes on billboards?
	Do you expect increasing risks with regard to future developments?
Economic, societal, and ethical impacts of Deepfakes	
➢ Economic impact	Do you think deepfake have an economic impact? How can the economic impact of Deepfakes be evaluated? What could be strategies to measure the economic consequences of deepfakes?
	Which branches next to (social) media will be affected, e.g. software-based communication, insurances, banking system, mobility (autonomous driving)?

	What kind of economic consequences could be expected, is a quantification possible? Which fields might be affected most? - stock market - public sector - branches - specific companies - individuals
	What are the consequences for safety or cybersecurity?
➢ Societal impact	Do you think Deepfakes have ethical implications? How do Deepfakes change our ethical framing / societal understanding of 'reality', 'truth', 'illusion' and 'Beguiling'? - Does this change societal perception in general? - What are the consequences for culture / evidence / epistemology in science or justice? - What are personal consequences (individual, psychological level)? How does this differ to 'fakes' in former times? - Do deepfakes pose new / unique challenges? - Maybe also as opposed to other fields / applications (Plagiarism; Photoshop) - What is unique about deepfakes?
	How do Deepfakes have a specific severe impact on minority groups, thus bear a special danger of discrimination? Exploitation potential women / children?
	Do you think Deepfakes challenges everyday life, education, and the sphere of employment? And if yes, how? • Which kind of information / awareness raising should be provided to society? • Are there further measures necessary, e.g. training for specific groups? • How can the awareness for Deepfakes embedded in school curricula? • Which new skills and qualifications for occupational groups arise, e.g. journalists, social media experts), rise of new occupations?
	Which role can culture & art play to start a societal reflection on deepfakes?
➢ Ethical / Political impact	Do Deepfakes influence democratic processes, e.g. political opinion forming and decision-making? In what way? How are societal values and fundamental rights challenged, e.g. free speech / free opinion vs. limitation of Deepfakes? What are the broader implications of disinformation with regard to violent conflicts and national safety? How could you envisage the consequences of disinformation if also true facts, videos or audios will be contested and deniable?
Conclusions	
➢ Outlook	From your perspective: are measures needed to mitigate the risks associated with deepfakes?

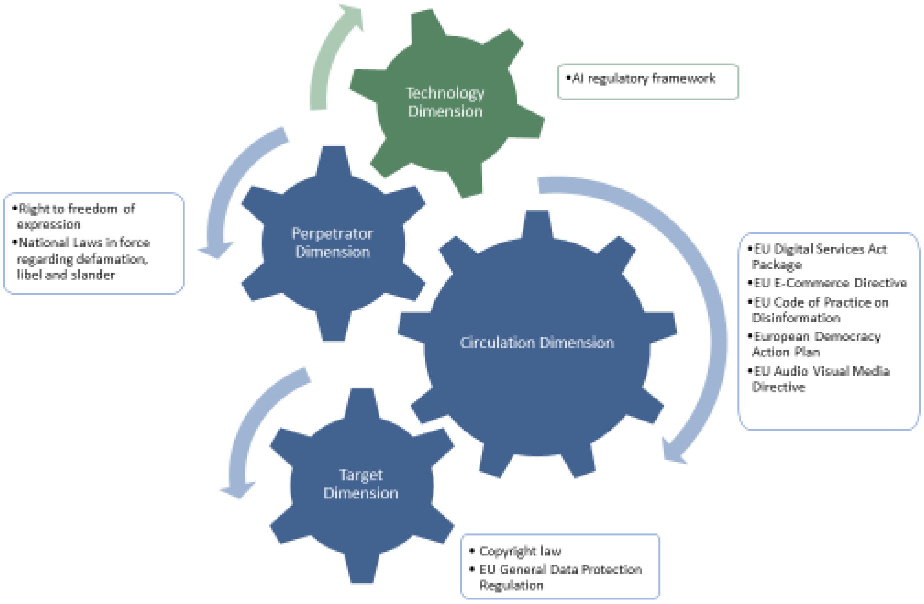
	Which kind of policy option is required most? How can this be governed best? By technical, social, or legal measures? Who should be responsible?
	What can we learn from previous attempts to try and mitigate adverse impacts of technological developments?
	Do you have further recommendations? Are there aspects not covered, yet?

Annex 3 – Interview questions second phase

Part	Questions
General	
- Introduction	Description of position, area of responsibility, professional and educational background; what is the relation to deep fakes
Key aspects for deepfake regulation	
- Legal understanding	What is your understanding of deepfakes – from a legal perspective?
- Concrete experiences	What are the most severe consequences/Impacts of deepfakes which should be regulated?
Existing regulations – most relevant policies at the European Level	
- Overview and Systematisation (Presentation of the figure which is send before the interview)	Is our picture complete?
	Do you know different approaches and dimensions of deepfake regulation?
	Do you miss additional dimensions and areas?
	Where would you identify regulatory gaps?
	Which dimensions are most relevant for the regulation of deepfakes?
- Our model in detail: Target Dimension Data protection, Copyright law	What is the relevance of this regulation for deepfakes? Could/ should it be strengthened/clarified?
- Our model in detail: - Circulation Dimension Digital platforms	What is the relevance of this regulation for deepfakes? Could/ should it be strengthened/clarified?
- Our model in detail: - Technology Dimension Production and detection of deepfakes by the new AI Framework	What is the relevance of this regulation for deepfakes? Could/ should it be strengthened/clarified?
	What opportunities do you see for the AI regulatory framework to regulate deepfakes?
	Do you know if deepfakes are considered in the AI framework?
	Why should they be considered?
	How could deepfakes be further addressed in the upcoming regulation?
	Do you agree with the risk-based approach for AI and deepfakes?

	Should deepfakes be considered as a 'high risk application'?
- Our model in detail: - Perpetrator dimension Cybercrime perspective	What is the relevance of this regulation for deepfakes? Could/ should it be strengthened/clarified?
Summing up measures of deepfake regulation	
- Gaps in existing law & regulation - Level of regulation	Are the challenges of deepfakes covered by existing EU regulations (national law)?
	How are the national and European policy levels intertwined when it comes to deepfakes?
	Is there an enforcement problem when it comes to deepfakes and if so, in what way?
- Integrative and multidimensional regulatory approach	What mix of existing and new regulations would be desirable to address the harmful consequences of deepfakes?
	What is the importance of the technology dimension/ AI regulatory framework in relation to others?
	Should there be a new integrative and multiperspective governance concept?
- Responsibility and actors	What regulatory models would you prefer (government regulation/ co-regulation/ self-regulation/other?)
	What is the role and responsibility of internet platforms in addressing the harmful consequences of deepfakes?
Conclusions	
- Outlook	From your perspective: Which kind of action from politics and other actors is required most?
- Summary	Do you have further recommendations? Are there aspects not covered, yet?

Our model in detail: The dimensions for the regulation of deepfakes


Technology Dimension
•AI regulatory framework
Perpetrator Dimension
•Right to freedom of expression
•National Laws in force regarding defamation, libel and slander
Circulation Dimension
•EU Digital Services Act Package
•EU E-Commerce Directive
•EU Code of Practice on Disinformation
•European Democracy Action Plan
•EU Audio Visual Media Directive
Target Dimension
• Copyright law
• EU General Data Protection Regulation

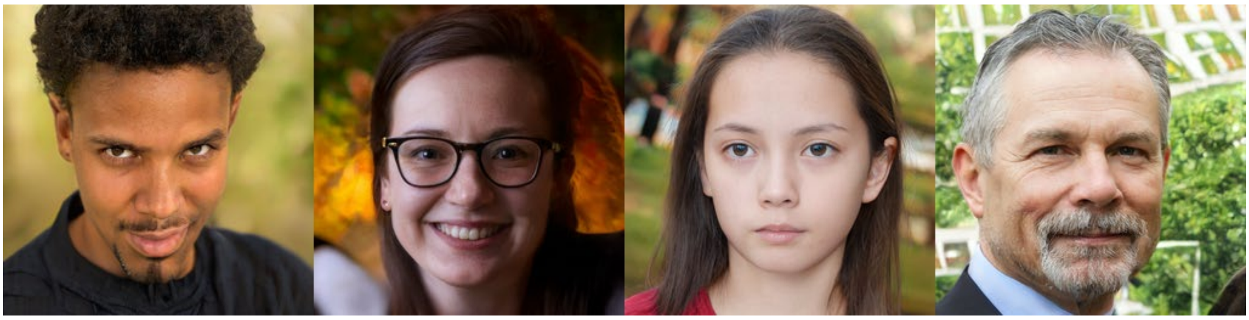
Annex 4 – GANs and Autoencoders

In this Annex two specific deepfake technologies are described in more detail: Generative Adversarial Networks and Autoencoders.

Generative Adversarial Networks

The above-mentioned developments - facial recognition, large datasets and image forensics - were foundational to the adoption by deepfake creators of a particular approach to Artificial Intelligence called Generative Adversarial Networks (GANs)[315]. In 2017, at the time of the emergence of deepfakes, scholars already produced over 16,000 papers mentioning GANs annually, indicating the common use of this technique.[316] In essence, GANs are computer programmes capable of generating a similar yet novel outcome compared to a training set by utilising a feedback loop learning strategy. The programmes consist of two competing elements. A so-called 'generative network' that creates content by analysing a large training dataset. In the case of deepfakes this generative network detects common patterns in pictures using facial recognition and creates similar content. Next, a 'discriminative network' that aims to identify forgeries based on forensics determines whether the created content is convincingly authentic or similar to the training set or not. Every time the discriminative network detects a forgery, the generative network takes note and tries to improve its outcome. GANs can be applied to create any kind of content, and over time dozens of GANs for generating specific visual content have been developed.[317] Recently, GANs became increasingly capable of generating human portrait pictures (See Figure 11).

Figure 11 - Four output images of a GAN capable of synthesising human portrait pictures that seem authentic[318]



Autoencoders

A second foundational technology is the development of so-called autoencoder programmes. Typical deepfake programmes such as face-swapping programmes (substituting the face of an individual in a target video footage with the face of a person in a source video) make use of this technique. It can be described in three steps. The first step is to detect and align the pose and facial expression in each and every frame of both a target and source video. Next, the programme learns how the facial features of a specific person change and relate to each other in particular expressions, such as simultaneous changes in a person mouth and eyebrows when smiling.

[315] Ian J. Goodfellow et al., 'Generative Adversarial Networks,' *ArXiv:1406.2661 [Cs, Stat]*, June 10, 2014.

[316] Kietzmann et al., 'Deepfakes.'

[317] Han Zhang et al., 'StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks,' *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 8 (August 2019): 1947–62; Christian Ledig et al., 'Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,' *ArXiv:1609.04802 [Cs, Stat]*, May 25, 2017.

[318] Karras et al., 'Analyzing and Improving the Image Quality of StyleGAN.'

The more images available, the better the programme gets at understanding these relationships. Once the training is complete, the programme is able to adjust the expression in any given picture of the target person by detecting the expression in a source picture.

REFERENCES

Agarwal, Sakshi, and Lav R. Varshney. 'Limits of Deepfake Detection: A Robust Estimation Viewpoint.' *ArXiv:1905.03493 [Cs, Math, Stat]*, May 9, 2019. http://arxiv.org/abs/1905.03493.

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 'Protecting World Leaders Against Deep Fakes,' 38–45, 2019. https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.html.

AI-Generated Facial Photos For 3D Human Creation | Headshot Plugin | Character Creator. 'AI-Generated Facial Photos For 3D Human Creation | Headshot Plugin | Character Creator.' Accessed May 11, 2021. https://www.reallusion.com/character-creator/headshot/ai-generated-face.html.

Ajder, Henry. 'The State of Deepfakes: Landscape, Threats, and Impact.' Sensity, 2019. https://sensity.ai/reports/.

Alattar, Adnan, Ravi Sharma, and John Scriven. 'A System for Mitigating the Problem of Deepfake News Videos Using Watermarking.' *Electronic Imaging* 2020, no. 4 (January 26, 2020): 117-1-117–10. https://doi.org/10.2352/ISSN.2470-1173.2020.4.MWSF-117.

Amerini, Irene, and Roberto Caldelli. 'Exploiting Prediction Error Inconsistencies through LSTM-Based Classifiers to Detect Deepfake Videos.' In *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, 97–102. IH&amp;MMSec '20. Denver, CO, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3369412.3395070.

Arik, Sercan O., Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, et al. 'Deep Voice: Real-Time Neural Text-to-Speech.' *ArXiv:1702.07825 [Cs]*, March 7, 2017. http://arxiv.org/abs/1702.07825.

Asarch, Steven. 'Wombo.Ai Lets Users Make Silly Deepfake Videos of Their Friends or Celebrities Singing Songs.' Insider, 2021. https://www.insider.com/wombo-ai-womboai-download-transforms-photo-a-singing-deepfake-face-2021-3.

'Assessment of the Code of Practice on Disinformation – Achievements and Areas for Further Improvement.' European Commission, 2020. https://digital-strategy.ec.europa.eu/en/library/assessment-code-practice-disinformation-achievements-and-areas-further-improvement.

Au, Lavender. 'China Targets 'deepfake' Content with New Regulation · TechNode.' TechNode, December 3, 2019. https://technode.com/2019/12/03/china-targets-deepfake-content-with-new-regulation/.

Ayyub, Rana. 'I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me.' HuffPost UK, November 21, 2018. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

Bao, Jianmin, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 'Towards Open-Set Identity Preserving Face Synthesis.' *ArXiv:1803.11182 [Cs]*, August 9, 2018. http://arxiv.org/abs/1803.11182.

Barari, S, C Lucas, and K Munger. 'Political Deepfake Videos Misinform the Public, But No More than Other Fake Media,' 2021. https://doi.org/10.31219/osf.io/cdfh3.

Bateman, Jon. 'Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenario.' Carnegie Endowment for International Peace, 2020. https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237.

Bayer, Judit, Natalija Bitiukova, Petra Bárd, Judit Szakács, Alberto Alemanno, and Erik Uszkiewicz. 'Disinformation and Propaganda – Impact on the Functioning of the Rule of Law in the EU and Its Member States - Think Tank.' Directorate General for Internal Policies of the Union, 2019. https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2019)608864.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? &#x1f99c;' In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Bennett, W Lance, and Steven Livingston. 'The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions.' *European Journal of Communication* 33, no. 2 (April 1, 2018): 122–39. https://doi.org/10.1177/0267323118760317.

Bitouk, Dmitri, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. 'Face Swapping: Automatically Replacing Faces in Photographs.' *ACM Transactions on Graphics* 27, no. 3 (August 1, 2008): 1–8. https://doi.org/10.1145/1360612.1360638.

Boheemen, Pieter van, Geert Munnichs, and Elma Dujso. 'Digital Threats to Democracy.' The Hague: Rathenau Instituut, 2020. https://www.rathenau.nl/en/digital-society/digital-threats-democracy.

Boheemen, Pieter van, Geert Munnichs, Linda Kool, Gijs Diercks, Jurriën Hamer, and Anouk Vos. 'Cyber Resilience with New Technology - An Opportunity and a Necessity.' Rathenau Instituut, 2020. https://www.rathenau.nl/sites/default/files/2020-07/REPORT%20Cyber%20resilience%20with%20new%20technology%20-%20Rathenau%20Instituut.pdf.

Bonfanti, Matteo. 'The Weaponisation of Synthetic Media: What Threat Does This Pose to National Security?' *CSS ETH Zurich* (blog), 2020. https://isnblog.ethz.ch/security/the-weaponisation-of-synthetic-media-what-threat-does-this-pose-to-national-security.

Botha, Johannes G., and Heloise Pieterse. *Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security*, 2020. https://doi.org/10.34190/ICCWS.20.085.

Boucher, Philip. 'Artificial Intelligence: How Does It Work, Why Does It Matter, and What Can We Do about It?' Panel for the Future of Science and Technology, 2020. https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2020)641547.

Bradshaw, S, and P Howard. 'Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation.' University of Oxford, 2018. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf.

Bregler, Christoph, Michelle Covelle, and Malcolm Slaney. 'Video Rewrite.' Interval Research Corporation, 1997. http://chris.bregler.com/videorewrite/.

Bressan, Sarah. 'Can the EU Prevent Deepfakes From Threatening Peace?' Carnegie Europe, 2019. https://carnegieeurope.eu/strategiceurope/79877.

Briscoe, Scott. 'U.S. Laws Address Deepfakes,' 2021. http://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 'Language Models Are Few-Shot Learners.' *ArXiv:2005.14165 [Cs]*, July 22, 2020. http://arxiv.org/abs/2005.14165.

Buzzfeed. 'You Won't Believe What Obama Says in This Video ☺.' Twitter, 2018. https://twitter.com/buzzfeed/status/986257991799222272.

Cahlan, Sarah. 'How Misinformation Helped Spark an Attempted Coup in Gabon.' *Washington Post*, 2020. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

Cao, Jie, Yibo Hu, Bing Yu, Ran He, and Zhenan Sun. '3D Aided Duet GANs for Multi-View Face Image Synthesis.' *IEEE Transactions on Information Forensics and Security* 14, no. 8 (August 2019): 2028–42. https://doi.org/10.1109/TIFS.2019.2891116.

Caporusso, Nicholas. 'Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology.' In *Advances in Artificial Intelligence, Software and Systems Engineering*, edited by Tareq Ahram, 1213:235–41. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-51328-3_33.

Carlini, Nicholas, and Hany Farid. 'Evading Deepfake-Image Detectors with White- and Black-Box Attacks.' *ArXiv:2004.00622 [Cs]*, April 1, 2020. http://arxiv.org/abs/2004.00622.

Chan, Caroline. *Everybody Dance Now*, 2018. https://www.youtube.com/watch?v=PCBTZh41Ris.

Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 'Everybody Dance Now.' *ArXiv:1808.07371 [Cs]*, August 27, 2019. http://arxiv.org/abs/1808.07371.

Chang, Xu, Jian Wu, Tongfeng Yang, and Guorui Feng. 'DeepFake Face Image Detection Based on Improved VGG Convolutional Neural Network.' In *2020 39th Chinese Control Conference (CCC)*, 7252–56, 2020. https://doi.org/10.23919/CCC50068.2020.9189596.

Chen, Mo, Jessica Fridrich, Miroslav Goljan, and Jan Lukas. 'Determining Image Origin and Integrity Using Sensor Noise.' *IEEE Transactions on Information Forensics and Security* 3, no. 1 (March 2008): 74–90. https://doi.org/10.1109/TIFS.2007.916285.

Cheng, Yi-Ting, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. '3D-Model-Based Face Replacement in Video.' In *SIGGRAPH '09: Posters*, 1. SIGGRAPH '09. New Orleans, Louisiana: Association for Computing Machinery, 2009. https://doi.org/10.1145/1599301.1599330.

Chesney, Robert, and Danielle Keats Citron. 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.' SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, July 14, 2018. https://papers.ssrn.com/abstract=3213954.

'China Regulators Held Talks with Alibaba, Tencent, Nine Others on 'deepfake' Tech.' *Reuters*, March 18, 2021. https://www.reuters.com/article/us-china-cyberspace-idUSKBN2BA09H.

Chintha, Akash, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. 'Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection.' *IEEE Journal of Selected Topics in Signal Processing* 14, no. 5 (August 2020): 1024–37. https://doi.org/10.1109/JSTSP.2020.2999185.

Chiu, Karen. 'China Announces New Rules to Tackle Deepfake Videos.' South China Morning Post, November 30, 2019. https://www.scmp.com/abacus/news-bites/article/3040033/china-announces-new-rules-tackle-deepfake-videos.

Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 'StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation.' *ArXiv:1711.09020 [Cs]*, September 21, 2018. http://arxiv.org/abs/1711.09020.

Ciancaglini, V, C Gibson, D Sancho, O McCarthy, M Eira, P Amann, and A Klayn. 'Malicious Uses and Abuses of Artificial Intelligence.' Trend Micro Research, United Nations Interregional Crime and Justice Research Institute & Europol's European Cybercrime Centre, November 19, 2020. https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence.

Clarke, Yvette D. 'Text - H.R.3230 - 116th Congress (2019-2020): Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019.' Legislation, June 28, 2019. https://www.congress.gov/bill/116th-congress/house-bill/3230/text.

Codec Avatars Side-by-Side Comparison. Tech@Facebook, 2019. https://www.facebook.com/TechAtFacebook/videos/codec-avatars-side-by-side-comparison/2383798615280431/.

Cole, Samantha. 'We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now,' 2018. https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

Collins, Aengus. 'Forged Authenticity: Governing Deepfake Risks,' 2019. https://doi.org/10.5075/EPFL-IRGC-273296.

Content Authenticity Initiative. 'Content Authenticity Initiative.' Accessed May 6, 2021. https://contentauthenticity.org.

Costa, Ana, Joost Bakker, and Gabriela Plucinska. 'How and Why It Works: The Principles and History behind Visual Communication.' *Medical Writing* 29 (March 1, 2020): 16–21.

Dale, Robert. 'GPT-3: What's It Good for?' *Natural Language Engineering* 27, no. 1 (2021): 113–18. https://doi.org/10.1017/S1351324920000601.

Damer, Naser, Alexandra Moseguí Saladié, Andreas Braun, and Arjan Kuijper. 'MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network.' In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–10, 2018. https://doi.org/10.1109/BTAS.2018.8698563.

De Keersmaecker, Jonas, and Arne Roets. 'Fake News': Incorrect, but Hard to Correct. The Role of Cognitive Ability on the Impact of False Information on Social Impressions.' *Intelligence* 65 (November 1, 2017): 107–10. https://doi.org/10.1016/j.intell.2017.10.005.

'De Toekomst van Online Platforms.' Rathenau Instituut, 2021. https://www.rathenau.nl/nl/berichten-aan-het-parlement/de-toekomst-van-online-platformen.

Deepfake. 'Faceswap.' *Github Repository*, 2021. https://github.com/deepfakes/faceswap.

'Deepfake Queen to Deliver Channel 4 Christmas Message.' *BBC News*, December 23, 2020, sec. Technology. https://www.bbc.com/news/technology-55424730.

Delfino, Rebecca. 'Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act.' *Fordham Law Review* 88, no. 3 (December 1, 2019): 887.

Deshmukh, Anushree, and Sunil B. Wankhade. 'Deepfake Detection Approaches Using Deep Learning: A Systematic Review.' In *Intelligent Computing and Networking*, edited by Valentina Emilia Balas, Vijay Bhaskar Semwal, Anand Khandare, and Megharani Patil, 293–302. Lecture Notes in Networks and Systems. Singapore: Springer, 2021. https://doi.org/10.1007/978-981-15-7421-4_27.

dfaker. 'Df.' *Github Repository*, 2021. https://github.com/dfaker/df.

Dietmar, Julia. 'GANs And Deepfakes Could Revolutionize The Fashion Industry.' Forbes, 2019. https://www.forbes.com/sites/forbestechcouncil/2019/05/21/gans-and-deepfakes-could-revolutionize-the-fashion-industry/.

Diresta, Renee. 'Free Speech Is Not the Same As Free Reach.' *Wired*, 2018. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 'Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?' *The International Journal of Press/Politics*, July 25, 2020, 1940161220944364. https://doi.org/10.1177/1940161220944364.

Du, Mengnan, Shiva Pentyala, Yuening Li, and Xia Hu. 'Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder.' *ArXiv:1909.05999 [Cs]*, September 19, 2020. http://arxiv.org/abs/1909.05999.

'EDMO – United against Disinformation.' Accessed May 20, 2021. https://edmo.eu/.

'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice.' ERGA, 2020. https://erga-online.eu/wp-content/uploads/2020/05/ERGA-2019-report-published-2020-LQ.pdf.

Ettema, Yori. 'Deepmemory.' Accessed May 11, 2021. https://yorie.nl/deepmemory/.

European Commission. 'Action Plan on Disinformation: Commission Contribution to the European Council,' 2018. https://ec.europa.eu/info/publications/action-plan-disinformation-commission-contribution-european-council-13-14-december-2018_en.

———. 'Code of Practice on Disinformation,' 2021. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

———. 'Communication on Tackling Online Disinformation: A European Approach,' 2018. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0236&qid=1621328030950.

———. 'Dual-Use Trade Controls,' 2018. https://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/.

———. 'European democracy action plan,' 2020. https://ec.europa.eu/info/strategy/priorities-2019-2024/new-push-european-democracy/european-democracy-action-plan_en.

———. 'GMO Legislation.' Text, October 17, 2016. https://ec.europa.eu/food/plant/gmo/legislation_en.

———. 'New Rules for Artificial Intelligence – Questions and Answers.' Text, 2021. https://ec.europa.eu/commission/presscorner/home/en.

———. 'Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (artificial intelligence act),' 2021. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence.

———. 'Security Union: European Commission Presents the Twentieth Progress Report.' Migration and Home Affairs - European Commission, October 30, 2019. https://ec.europa.eu/home-affairs/news/20191030_security-union-european-commission-presents-twentieth-progress-report_en.

———. 'The digital services act Package,' 2021. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

European Parliament, „Resolution on online platforms and the digital single market'. P8_TA(2017)0272, June 15th 2017.

———.„Resolution of 3 May 2018 on media pluralism and media freedom in the European Union', P8_TA(2018)0204, May 3th 2018.

———.„Resolution of 12 February 2019 on a comprehensive European industrial policy on artificial intelligence and robotics'. P8_TA(2019)0081, February 12th 2019, Nr. 178.

———.„Recommendation of 13 March 2019 to the Council and the Vice- President of the Commission / High Representative of the Union for Foreign Affairs and Security Policy concerning taking stock of the follow-up taken by the EEAS two years after the EP report on EU strategic communication to counteract propaganda against it by third parties'. P8_TA(2019)0187, March 13th 2019.

———. LIBE Committee. „Opinion of the Committee on Civil Liberties, Justice and Home Affairs for the Committee on Legal Affairs with recommendations to the Commission on the framework of ethical aspects of artificial intelligence, robotics and related technologies'. PE652.296v02-00, September 22nd 2020.

———.„Report on intellectual property rights for the development of artificial intelligence technologies.' P9_TA (2020)0277 October 20th 2020

———. JURI. „Report on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice'. PE653.860v02-00, January 4th 2021.

———. „Resolution of 19 May 2021 on artificial intelligence in education, culture and the audiovisual sector'. P9_TA(2021)0238 May 19th 2021.

Facebook Technology. 'Facebook Is Building the Future of Connection with Lifelike Avatars,' March 13, 2019. https://tech.fb.com/codec-avatars-facebook-reality-labs/.

Farid, Hany. 'Hany Farid: Deepfakes Give New Meaning to the Concept of 'fake News,' and They're Here to Stay.' Text.Article. Fox News, June 16, 2019. https://www.foxnews.com/opinion/hany-farid-deep-fakes.

Feeney, Matthew. 'Deepfake Laws Risk Creating More Problems Than They Solve.' Regulatory Transparency Project, 2021. https://regproject.org/paper/deepfake-laws-risk-creating-more-problems-than-they-solve/.

Ferraro, Matthew, and Louis Tompros. 'New York's Right to Publicity and Deepfakes Law Breaks New Ground,' 2020. https://www.wilmerhale.com/en/insights/client-alerts/20201217-new-yorks-right-to-publicity-and-deepfakes-law-breaks-new-ground.

Flynn, D. J., Brendan Nyhan, and Jason Reifler. 'The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics.' *Political Psychology* 38, no. S1 (2017): 127–50. https://doi.org/10.1111/pops.12394.

Galston, William A. 'Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics.' *Brookings* (blog), January 8, 2020. https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

Galvan, Fausto, and Arma Carabinieri. 'Image/Video Forensics.' CEPOL, 2020. European Law Enforcement Research Bulletin. https://bulletin.cepol.europa.eu/index.php/bulletin/article/view/399.

Gensing, Patrick. *Fakten gegen Fake News oder Der Kampf um die Demokratie*. Duden, 2019.

Gong, Dafeng. 'Deepfake Forensics, an AI-Synthesized Detection with Deep Convolutional Generative Adversarial Networks.' *International Journal of Advanced Trends in Computer Science and Engineering* 9, no. 3 (June 25, 2020): 2861–70. https://doi.org/10.30534/ijatcse/2020/58932020.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 'Generative Adversarial Networks.' *ArXiv:1406.2661 [Cs, Stat]*, June 10, 2014. http://arxiv.org/abs/1406.2661.

Goodfellow, Ian, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel, and Jack Clark. 'Attacking Machine Learning with Adversarial Examples.' OpenAI, February 24, 2017. https://openai.com/blog/adversarial-example-research/.

Greene, Viveca S. 'Deplorable' Satire: Alt-Right Memes, White Genocide Tweets, and Redpilling Normies.' *Studies in American Humor* 5, no. 1 (2019): 31–69. https://doi.org/10.5325/studamerhumor.5.1.0031.

Groh, Matt. 'Project Overview ‹ Detect DeepFakes: How to Counteract Misinformation Created by AI.' MIT Media Lab, 2020. https://www.media.mit.edu/projects/detect-fakes/overview/.

Hamer, Jurriën, Rinie van Est, and Lambèr Royakkers. 'Cyberspace without Conflict.' Rathenau Instituuut, 2019. https://www.rathenau.nl/en/digital-society/cyberspace-without-conflict.

Hao, Karen. 'We Read the Paper That Forced Timnit Gebru out of Google. Here's What It Says.' MIT Technology Review, 2020. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

Harris, Douglas. 'Deepfakes: False Pornography Is Here and the Law Cannot Protect You.' *Duke Law & Technology Review* 17, no. 1 (January 5, 2019): 99–127.

Hartmann, K., and K. Giles. 'The Next Generation of Cyber-Enabled Information Warfare.' In *2020 12th International Conference on Cyber Conflict (CyCon)*, 1300:233–50, 2020. https://doi.org/10.23919/CyCon49761.2020.9131716.

Hasan, Haya R., and Khaled Salah. 'Combating Deepfake Videos Using Blockchain and Smart Contracts.' *IEEE Access* 7 (2019): 41596–606. https://doi.org/10.1109/ACCESS.2019.2905689.

Haysom, Sam. 'People Are Using Face-Swapping Tech to Add Nicolas Cage to Random Movies and What Is 2018.' Mashable, 2018. https://mashable.com/2018/01/31/nicolas-cage-face-swapping-deepfakes/.

He, Zhenliang, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 'AttGAN: Facial Attribute Editing by Only Changing What You Want.' *ArXiv:1711.10678 [Cs, Stat]*, July 25, 2018. http://arxiv.org/abs/1711.10678.

Hewage, Chaminda. 'Data Protection in the Wake of Deepfakes.' Infosecurity Magazine, 2020. https://www.infosecurity-magazine.com:443/next-gen-infosec/data-protection-wake-deepfakes/.

High Representative of the Union for Foreign Affairs and Security Policy. 'Report on the Implementation of the Action Plan Against Disinformation.' Brussels, 2019. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019JC0012&rid=8.

Hongmeng, Zhang, Zhu Zhiqiang, Sun Lei, Mao Xiuqing, and Wang Yuehan. 'A Detection Method for DeepFake Hard Compressed Videos Based on Super-Resolution Reconstruction Using CNN.' In *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, 98–103. HPCCT &amp; BDAI 2020. Qingdao, China: Association for Computing Machinery, 2020. https://doi.org/10.1145/3409501.3409542.

'How Facebook Can Flatten the Curve of the Coronavirus Infodemic.' Avaaz, 2020. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/.

'How Powerful Are Social Bots?' Günther Thiele Foundation, 2018. https://www.akademische-gesellschaft.com/fileadmin/webcontent/Publikationen/Communication_Snapshots/AGUK_CommunicationSnapshot_SocialBots_June2018.pdf.

'How to Use AI Generated Photos | Generated.Photos.' Accessed May 4, 2021. https://generated.photos/use-cases.

Huang, Yihao, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. 'FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction.' *ArXiv:2006.07533 [Cs]*, August 17, 2020. http://arxiv.org/abs/2006.07533.

Hussain, Shehzeen, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. 'Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples.' *ArXiv:2002.12749 [Cs]*, November 7, 2020. http://arxiv.org/abs/2002.12749.

IBM. 'What Is Computer Vision?' Accessed March 22, 2021. https://www.ibm.com/topics/computer-vision.

'Internet Organised Crime Threat Assessment (IOCTA) 2019.' Europol, 2019. https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019.

iperov. 'DeepFaceLab.' *Github Repository*, 2021. https://github.com/iperov/DeepFaceLab.

Iqbal, Mansoor. 'FaceApp Revenue and Usage Statistics (2020).' Business of Apps, September 5, 2019. https://www.businessofapps.com/data/faceapp-statistics/.

Jafar, Mousa Tayseer, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. 'Forensics and Analysis of Deepfake Videos.' In *2020 11th International Conference on Information and Communication Systems (ICICS)*, 053–058, 2020. https://doi.org/10.1109/ICICS49469.2020.239493.

Jain, Simran, and Piyush Jha. 'Deepfakes in India: Regulation and Privacy.' *South Asia@LSE* (blog), May 21, 2020. https://blogs.lse.ac.uk/southasia/2020/05/21/deepfakes-in-india-regulation-and-privacy/.

Jiacheng, Shang, Si Chen, and Jie Wu. 'Defending Against Voice Spoofing: A Robust Software-Based Liveness Detection System.' In *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. Chengdu: IEEE, 2018. https://doi.org/10.1109/MASS.2018.00016.

Jokubauskas, Remigijus, and Marek Świerczyński. 'Is Revision of the Council of Europe Guidelines on Electronic Evidence Already Needed?' *Utrecht Law Review* 16, no. 1 (May 26, 2020): 13–20. https://doi.org/10.36633/ulr.525.

Jung, Tackhyun, Sangwon Kim, and Keecheon Kim. 'DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern.' *IEEE Access* 8 (2020): 83144–54. https://doi.org/10.1109/ACCESS.2020.2988660.

Kalf, Sanne. 'What Does a Feminist Approach to Deepfake Pornography Look Like?,' October 24, 2019. http://mastersofmedia.hum.uva.nl/blog/2019/10/24/what-does-a-feminist-approach-to-deepfake-pornography-look-like/.

Kalpokas, Ignas. 'Problematising Reality: The Promises and Perils of Synthetic Media.' *SN Social Sciences* 1, no. 1 (November 9, 2020): 1. https://doi.org/10.1007/s43545-020-00010-8.

Kamble, Madhu R., Hardik B. Sailor, Hemant A. Patil, and Haizhou Li. 'Advances in Anti-Spoofing: From the Perspective of ASVspoof Challenges.' *APSIPA Transactions on Signal and Information Processing* 9 (ed 2020). https://doi.org/10.1017/ATSIP.2019.21.

Karras, Tero, Samuli Laine, and Timo Aila. 'A Style-Based Generator Architecture for Generative Adversarial Networks.' *ArXiv:1812.04948 [Cs, Stat]*, March 29, 2019. http://arxiv.org/abs/1812.04948.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 'Analyzing and Improving the Image Quality of StyleGAN.' *ArXiv:1912.04958 [Cs, Eess, Stat]*, March 23, 2020. http://arxiv.org/abs/1912.04958.

Ker, Nic. 'Is the Political Aide Viral Sex Video Confession Real or a Deepfake? | Malay Mail,' 2019. https://www.malaymail.com/news/malaysia/2019/06/12/is-the-political-aide-viral-sex-video-confession-real-or-a-deepfake/1761422.

Kerner, Catherine, and Mathias Risse. 'Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds.' *Moral Philosophy and Politics*, November 11, 2020. https://doi.org/10.1515/mopp-2020-0024.

Khodabakhsh, Ali, and Christoph Busch. 'A Generalizable Deepfake Detector Based on Neural Conditional Distribution Modelling.' In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–5, 2020.

Kietzmann, Jan, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 'Deepfakes: Trick or Treat?' *Business Horizons*, ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING, 63, no. 2 (March 1, 2020): 135–46. https://doi.org/10.1016/j.bushor.2019.11.006.

Kim, Hyeongwoo, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. 'Neural Style-Preserving Visual Dubbing.' *ACM Transactions on Graphics* 38, no. 6 (November 8, 2019): 1–13. https://doi.org/10.1145/3355089.3356500.

Kirchengast, Tyrone. 'Deepfakes and Image Manipulation: Criminalisation and Control.' *Information & Communications Technology Law* 29, no. 3 (September 1, 2020): 308–23. https://doi.org/10.1080/13600834.2020.1794615.

Korshunova, Iryna, Wenzhe Shi, Joni Dambre, and Lucas Theis. 'Fast Face-Swap Using Convolutional Neural Networks.' *ArXiv:1611.09577 [Cs]*, July 27, 2017. http://arxiv.org/abs/1611.09577.

Kwok, Andrei O. J., and Sharon G. M. Koh. 'Deepfake: A Social Construction of Technology Perspective.' *Current Issues in Tourism* 0, no. 0 (March 14, 2020): 1–5. https://doi.org/10.1080/13683500.2020.1738357.

Lantwin, Tobias. 'Deep Fakes – Düstere Zeiten Für Den Persönlichkeitsschutz? Rechtliche Herausforderungen Und Lösungsansätze.' *MultiMedia Und Recht*, 2019. https://www.dropbox.com/s/3wjzywxvk1ow8o6/MMR%2009-2019%20-%20Beitrag%20Lantwin.pdf?dl=0.

Ledig, Christian, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, et al. 'Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network.' *ArXiv:1609.04802 [Cs, Stat]*, May 25, 2017. http://arxiv.org/abs/1609.04802.

Lewandowsky, Stephan, Laura Smillie, David Garcia, Ralph Hertwig, Jim Weatherall, Stephanie Egidy, Ronald Robertson, et al. 'Technology and Democracy: Understanding the Influence of Online Technologies on Political Behaviour and Decision-Making.' JRC, 2020. https://publications.jrc.ec.europa.eu/repository/handle/JRC122023.

Lexico Dictionaries. 'Avatar.' Accessed March 22, 2021. https://www.lexico.com/definition/avatar.

Li, Lingzhi, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 'FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping.' *ArXiv:1912.13457 [Cs]*, September 15, 2020. http://arxiv.org/abs/1912.13457.

Li, Lingzhi, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 'Face X-Ray for More General Face Forgery Detection.' *ArXiv:1912.13458 [Cs]*, April 18, 2020. http://arxiv.org/abs/1912.13458.

Li, Yuezun, Ming-Ching Chang, and Siwei Lyu. 'In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking.' *ArXiv:1806.02877 [Cs]*, June 11, 2018. http://arxiv.org/abs/1806.02877.

Li, Yuezun, and Siwei Lyu. 'Exposing DeepFake Videos By Detecting Face Warping Artifacts.' *ArXiv:1811.00656 [Cs]*, May 22, 2019. http://arxiv.org/abs/1811.00656.

Li, Yuezun, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 'Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics.' *ArXiv:1909.12962 [Cs, Eess]*, March 16, 2020. http://arxiv.org/abs/1909.12962.

Lyu, Siwei. 'DeepFake Detection: Current Challenges and Next Steps.' *ArXiv:2003.09234 [Cs]*, March 11, 2020. http://arxiv.org/abs/2003.09234.

Maddocks, Sophie. 'A Deepfake Porn Plot Intended to Silence Me': Exploring Continuities between Pornographic and 'Political' Deep Fakes.' *Porn Studies* 7, no. 4 (October 1, 2020): 415–23. https://doi.org/10.1080/23268743.2020.1757499.

Madiega, Tambiama. 'Reform of the EU Liability Regime for Online Intermediaries : Background on the Forthcoming digital services act : In-Depth Analysis.' Website. European Parliament, May 12,

2020. http://op.europa.eu/en/publication-detail/-/publication/81cddd82-94c2-11ea-aac4-01aa75ed71a1/language-en.

Malik, Hafiz, and Raghavendar Changalvala. 'Fighting AI with AI: Fake Speech Detection Using Deep Learning.' Audio Engineering Society, 2019. https://www.aes.org/e-lib/browse.cfm?elib=20479.

Martin, Kim, and V. P. Marketing. 'What Is AI Voice Cloning Software? Find Out at ID R&D.' *ID R&D* (blog), March 9, 2020. https://www.idrnd.ai/what-is-voice-cloning/.

McCoy, Erin. 'Visual Communication Is Transforming Marketing -- Are You Up To Speed?' Forbes, 2017. https://www.forbes.com/sites/forbescommunicationscouncil/2017/05/12/visual-communication-is-transforming-marketing-are-you-up-to-speed/.

McPeak, Agnieszka. 'The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood.' SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, February 21, 2021. https://papers.ssrn.com/abstract=3824021.

Merriam Webster. 'Words We're Watching: 'Deepfake.' Accessed April 28, 2021. https://www.merriam-webster.com/words-at-play/deepfake-slang-definition-examples.

Meskys, Edvinas, Aidas Liaudanskas, Julija Kalpokiene, and Paulius Jurcys. 'Regulating Deep Fakes: Legal and Ethical Considerations.' *Journal of Intellectual Property Law & Practice* 15, no. 1 (January 1, 2020): 24–31. https://doi.org/10.1093/jiplp/jpz167.

Metz, Cade, and Keith Collins. 'How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos.' *The New York Times*, January 2, 2018, sec. Technology. https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html, https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. 'Kamerbrief over Beleidsinzet Bescherming Democratie Tegen Desinformatie,' 2019. www.rijksoverheid.nl/documenten/kamerstukken/2019/10/18/kamerbrief-overbeleidsinzet-bescherming-democratie-tegen-desinformatie.

Minsky, Carly. 'Deepfake' Videos: To Believe or Not Believe?,' January 26, 2021. https://www.ft.com/content/803767b7-2076-41e2-a587-1f13c77d1675.

Nagano, Koki, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. 'PaGAN: Real-Time Avatars Using Dynamic Textures.' *ACM Transactions on Graphics* 37, no. 6 (December 4, 2018): 258:1-258:12. https://doi.org/10.1145/3272127.3275075.

Naika, Ravika. 'An Overview of Automatic Speaker Verification System.' In *Intelligent Computing and Information and Communication*, edited by Subhash Bhalla, Vikrant Bhateja, Anjali A. Chandavale, Anil S. Hiwale, and Suresh Chandra Satapathy, 603–10. Advances in Intelligent Systems and Computing. Singapore: Springer, 2018. https://doi.org/10.1007/978-981-10-7245-1_59.

Nash, Jim. 'Bias in Facial Recognition Is Handicapping Deepfake Detection.' Biometric Update, May 17, 2021. https://www.biometricupdate.com/202105/bias-in-facial-recognition-is-handicapping-deepfake-detection.

Neelima, Medikonda, and I Santiprabha. 'Mimicry Voice Detection Using Convolutional Neural Networks.' In *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 314–18. Trichy, India: IEEE, 2020. https://doi.org/10.1109/ICOSEC49089.2020.9215407.

Nema, Purvi. 'Are Indian Laws Equipped To Deal With Deepfakes?' *The Journal of Indian Law and Society Blog* (blog), July 19, 2020. https://jilsblognujs.wordpress.com/2020/07/19/are-indian-laws-equipped-to-deal-with-deepfakes/.

Newman, Lily Hay. 'Police Bodycams Can Be Hacked to Doctor Footage.' *Wired*, 2018. https://www.wired.com/story/police-body-camera-vulnerabilities/.

Nguyen, Hoang Mark, and Reza Derakhshani. 'Eyebrow Recognition for Identifying Deepfake Videos.' In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–5, 2020.

Nguyen, Thanh Thi, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 'Deep Learning for Deepfakes Creation and Detection: A Survey.' *ArXiv:1909.11573 [Cs, Eess]*, July 28, 2020. http://arxiv.org/abs/1909.11573.

NOS. 'Politiek en meldpunt binden strijd aan met 'deep nudes,' 2020. https://nos.nl/l/2358075.

Niessner, Matthias. *Face2Face: Real-Time Face Capture and Reenactment of RGB Videos (CVPR 2016 Oral)*, 2016. https://www.youtube.com/watch?v=ohmajJTcpNk.

Nirkin, Yuval, Yosi Keller, and Tal Hassner. 'FSGAN: Subject Agnostic Face Swapping and Reenactment.' *ArXiv:1908.05932 [Cs]*, August 16, 2019. http://arxiv.org/abs/1908.05932.

Nirkin, Yuval, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gerard Medioni. 'On Face Segmentation, Face Swapping, and Face Perception.' *ArXiv:1704.06729 [Cs]*, April 21, 2017. http://arxiv.org/abs/1704.06729.

NVIDIA Developer. 'NVIDIA Maxine Video Conferencing Platform,' October 1, 2020. https://developer.nvidia.com/maxine.

Oord, Aäron van den, and Sander Dieleman. 'WaveNet: A Generative Model for Raw Audio.' Deepmind. Accessed January 25, 2021. https://deepmind.com/blog/wavenet-generative-model-raw-audio/.

Oxford Reference. 'Artificial Intelligence.' Accessed March 22, 2021. https://doi.org/10.1093/oi/authority.20110803095426960.

Paris, Britt, and Joan Donovan. 'Deepfakes and Cheap Fakes.' Data & Society, September 18, 2019. https://datasociety.net/library/deepfakes-and-cheap-fakes/.

Parker, Trey, Matt Stone, and Peter Serafinowicz. *Sassy Justice with Fred Sassy (Full Episode)*, 2020. https://www.youtube.com/watch?v=9WfZuNceFDM.

Patrini, Georgio. 'Mapping the Deepfake Landscape.' *Sensity* (blog), October 7, 2019. https://sensity.ai/mapping-the-deepfake-landscape/.

Patrini, Giorgio. 'Automating Image Abuse: Deepfake Bots on Telegram.' *Sensity* (blog), October 20, 2020. https://sensity.ai/automating-image-abuse-deepfake-bots-on-telegram/.

Pawelec, Maria, Cora Bieß, and Alexander Orlowski. 'Ethisch Und Sozial Wünschenswerte Technikgovernance Fördern.' Eberhard Karls Universität Tübingen, 2021. https://uni-tuebingen.de/einrichtungen/zentrale-einrichtungen/internationales-zentrum-fuer-ethik-in-den-wissenschaften/das-izew/newsfullview-aktuelles/article/ethisch-und-sozial-wuenschenswerte-technikgovernance-foerdern.

Petitions - UK Government and Parliament. 'Petition: Criminalise Manufacturing and Distributing Deep-Fake Pornography,' 2021. https://petition.parliament.uk/petitions/567793.

Pfefferkorn, Riana. 'Too Good to Be True? 'Deepfakes' Pose a New Challenge for Trial Courts.' *Washington State Bar Association*, 2019. https://law.stanford.edu/publications/too-good-to-be-true-deepfakes-pose-a-new-challenge-for-trial-courts/.

Pinscreen. 'Unreal PaGAN: AI-Generated Real-Time Avatars in UE4,' 2020. https://www.pinscreen.com/unrealpagan/.

'Pinscreen Virtual Assistant (Live Demo 2020) - YouTube.' Accessed February 15, 2021. https://www.youtube.com/watch?v=8MjhIQZt76c.

Plasilova, Iva, Jordan Hill, Marlin Carlberg, Marion Goubet, and Richard Procee. 'Study for the Assessment of the Implementation of the Code of Practice on Disinformation,' 2020. https://digital-strategy.ec.europa.eu/en/library/study-assessment-implementation-code-practice-disinformation.

Poelmans, Kyrill. 'What Is Natural Language Processing (NLP)?' Textmetrics, June 25, 2020. https://www.textmetrics.com/what-is-natural-language-processing-nlp.

Pollicino, Oreste, Giovanni De Gregorio, and Laura Somaini. 'The European Regulatory Conundrum to Face the Rise and Amplification of False Content Online.' SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 2020. https://papers.ssrn.com/abstract=3725528.

Radmilovic, Ines K., Tamás Bereczki, and Ádám Liber. 'Hungary Adopts National GDPR Supplementing Legislation,' 2018. https://www.lexology.com/library/detail.aspx?g=3e045c27-67c6-43a5-a5e8-a0a30a07612b.

Rasser, Martijn. 'Why Are Deepfakes So Effective?' Scientific American Blog Network, 2019. https://blogs.scientificamerican.com/observations/why-are-deepfakes-so-effective/.

Rathenau Instituut. 'Digitalisering van het nieuws: online nieuwsgedrag en personalisatie in Nederland.,' 2018. https://www.rathenau.nl/nl/digitale-samenleving/digitalisering-van-het-nieuws.

Rathenau Instituut. 'Rathenau Manifesto: Set 10 design requirements for tomorrow's digital society now,' 2020. https://www.rathenau.nl/en/manifest.

Reda, Julia. 'Der digital services act steht für einen Sinneswandel in Brüssel.' *Netzpolitik.org* (blog), 2021. https://netzpolitik.org/2021/edit-policy-der-digital-services-act-steht-fuer-einen-sinneswandel-in-bruessel/.

Reynolds, Matt. 'Courts and Lawyers Struggle with Growing Prevalence of Deepfakes.' ABA Journal, 2020. https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes.

'Right to the Protection of One's Image.' *European Court of Human Rights*, 2020. https://www.echr.coe.int/documents/fs_own_image_eng.pdf.

Rini, Regina. 'Deepfakes and the Epistemic Backstop.' *Philosopher's Imprint* 20, no. 24 (2020). http://hdl.handle.net/2027/spo.3521354.0020.024.

Roozenbeek, Jon, Sander van der Linden, and Thomas Nygren. 'Prebunking Interventions Based on 'Inoculation' Theory Can Reduce Susceptibility to Misinformation across Cultures.' *Harvard Kennedy School Misinformation Review* 1, no. 2 (February 3, 2020). https://doi.org/10.37016//mr-2020-008.

Rössler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 'FaceForensics++: Learning to Detect Manipulated Facial Images.' *ArXiv:1901.08971 [Cs]*, August 26, 2019. http://arxiv.org/abs/1901.08971.

Roth, Andrew. 'European MPs Targeted by Deepfake Video Calls Imitating Russian Opposition.' *The Guardian*, 2021. https://www.theguardian.com/world/2021/apr/22/european-mps-targeted-by-deepfake-video-calls-imitating-russian-opposition.

Runway. 'Runway | Make the Impossible.' Accessed May 4, 2021. https://www.runwayml.com.

Sagar, Ram. 'OpenAI Releases GPT-3, The Largest Model So Far.' *Analytics India Magazine* (blog), June 3, 2020. https://analyticsindiamag.com/open-ai-gpt-3-language-model/.

Salvador Dalí Museum. 'Dalí Lives (via Artificial Intelligence).' Accessed February 17, 2021. https://thedali.org/exhibit/dali-lives/.

Sasse, Ben. 'S.3805 - 115th Congress (2017-2018): Malicious Deep Fake Prohibition Act of 2018.' Webpage, December 21, 2018. https://www.congress.gov/bill/115th-congress/senate-bill/3805.

Schapiro, Zachary. 'DEEP FAKES Accountability Act: Overbroad and Ineffective – Intellectual Property and Technology Forum,' 2020. http://bciptf.org/2020/04/deepfakes-accountability-act/.

Scherhag, Ulrich, Luca Debiasi, Christian Rathgeb, Christoph Busch, and Andreas Uhl. 'Detection of Face Morphing Attacks Based on PRNU Analysis.' *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019. https://doi.org/10.1109/TBIOM.2019.2942395.

Schick, Nina. *Deep Fakes and the Infocalypse*. Octopus Publishing Group, 2020.

Sciforce. 'Text-to-Speech Synthesis: An Overview.' Medium, February 13, 2020. https://medium.com/sciforce/text-to-speech-synthesis-an-overview-641c18fcd35f.

Šepec, Miha, and Melanija Lango. 'Virtual Revenge Pornography as a New Online Threat to Sexual Integrity.' *Balkan Social Science Review* 15 (June 25, 2020): 117–35. https://doi.org/10.46763/BSSR20150118sh.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, et al. 'Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions.' *ArXiv:1712.05884 [Cs]*, February 15, 2018. http://arxiv.org/abs/1712.05884.

Smith, Nicola. 'Taiwan Builds 'nerd Immunity' to Resist Chinese Disinformation Campaigns.' *The Telegraph*, June 13, 2020. https://www.telegraph.co.uk/news/2020/06/13/taiwan-builds-nerd-immunity-resist-chinese-disinformation-campaigns/.

'Snapshot Paper - Deepfakes and Audiovisual Disinformation.' Centre for Data Ethics and Innovation, 2019. https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation.

Snijders, Dhoya, Sophie Horsman, Linda Kool, and Rinie van Est. 'Responsible VR. Protect Consumers in Virtual Reality.' Rathenau Instituut, 2020. https://www.rathenau.nl/sites/default/files/2020-03/Responsible%20VR.pdf.

Sokolov, S. S., O. M. Alimov, D. A. Tyapkin, Y. F. Katorin, and A. I. Moiseev. 'Modern Social Engineering Voice Cloning Technologies.' In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 513–16, 2020. https://doi.org/10.1109/EIConRus49466.2020.9038954.

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 'Synthesizing Obama: Learning Lip Sync from Audio.' *ACM Transactions on Graphics* 36, no. 4 (July 20, 2017): 95:1-95:13. https://doi.org/10.1145/3072959.3073640.

Synodinou, Tatiana. 'Image Right and Copyright Law in Europe: Divergences and Convergences.' *Laws* 3, no. 2 (2014): 181–207. https://doi.org/10.3390/laws3020181.

'Synthesia Insights: Case Study - David Beckham / Malaria No More / RGA,' 2020. https://www.synthesia.io/post/case-study-david-beckham-malaria-no-more-rga.

Taigman, Yaniv, Lior Wolf, Adam Polyak, and Eliya Nachmani. 'VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop.' *ArXiv:1707.06588 [Cs]*, February 1, 2018. http://arxiv.org/abs/1707.06588.

Ternovski, John, Joshua Kalla, and Peter Aronow. 'Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments.' *OSF Preprints*, 2021. https://doi.org/10.31219/osf.io/dta97.

Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 'Face2Face: Real-Time Face Capture and Reenactment of RGB Videos.' *Computer Vision Foundation*, 2016. https://web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/paper.pdf.

Tolosana, Ruben, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 'DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection.' *ArXiv:2001.00179 [Cs]*, June 18, 2020. http://arxiv.org/abs/2001.00179.

'Truepic | Photo and Video Verification Platform.' Accessed November 26, 2020. https://truepic.com/.

Tursman, Eleanor, Marilyn George, Seny Kamara, and James Tompkin. 'Towards Untrusted Social Video Verification to Combat Deepfakes via Face Geometry Consistency.' In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2784–93, 2020. https://doi.org/10.1109/CVPRW50498.2020.00335.

Vaccari, Cristian, and Andrew Chadwick. 'Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News.' *Social Media + Society* 6, no. 1 (January 1, 2020): 2056305120903408. https://doi.org/10.1177/2056305120903408.

Venema, Agnes, and Zeno Geradts. 'Digital Forensics, Deepfakes, and the Legal Process.' *The SciTech Lawyer*, 2020. https://www.americanbar.org/groups/science_technology/publications/scitech_lawyer/2020/summer/digital-forensics-deepfakes-and-legal-process/.

Vincent, James. 'Deepfake' That Supposedly Fooled European Politicians Was Just a Look-Alike, Say Pranksters,' 2021. https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus.

Wang, Zejian, Koki Nagano, Hao Li, Liwen Hu, Lain Goldwhite, Hanwei Kung, Aviral Agarwal, et al. 'AI-Synthesized Avatars: From Real-Time Deepfakes to Virtual AI Companions.' In *ACM SIGGRAPH 2020 Real-Time Live!*, 1. SIGGRAPH '20. Virtual Event, USA: Association for Computing Machinery, 2020. https://doi.org/10.1145/3407662.3415279.

Wardle, Claire, and Media Derakhshan. 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking.' *Shorenstein Center*, October 31, 2017.

https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/.

Washington Post. *Pelosi Videos Manipulated to Make Her Appear Drunk Are Being Shared on Social Media*, 2019. https://www.youtube.com/watch?v=sDOo5nDJwgA.

Weiten, Wayne. *Psychology: Themes and Variations: Themes and Variations*. Wadsworth/Cengage Learning, 2010.

Whittaker, L., T. C. Kietzmann, J. Kietzmann, and A. Dabirian. 'All around Me Are Synthetic Faces': The Mad World of AI-Generated Media.' *99*, 2020. https://repository.ubn.ru.nl/handle/2066/222433.

Wu, Jian, Kai Feng, Xu Chang, and Tongfeng Yang. 'A Forensic Method for DeepFake Image Based on Face Recognition.' In *Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, 104–8. HPCCT &amp; BDAI 2020. Qingdao, China: Association for Computing Machinery, 2020. https://doi.org/10.1145/3409501.3409544.

Xia, Yiping, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. 'Disinformation, Performed: Self-Presentation of a Russian IRA Account on Twitter.' *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1646–64. https://doi.org/10.1080/1369118X.2019.1621921.

Yadlin-Segal, Aya, and Yael Oppenheim. 'Whose Dystopia Is It Anyway? Deepfakes and Social Media Regulation.' *Convergence* 27, no. 1 (February 1, 2021): 36–51. https://doi.org/10.1177/1354856520923963.

Yang, Xin, Yuezun Li, and Siwei Lyu. 'Exposing Deep Fakes Using Inconsistent Head Poses.' *ArXiv:1811.00661 [Cs]*, November 13, 2018. http://arxiv.org/abs/1811.00661.

Yang, Zehi. 'Chinese Deepfakes Are Going Viral, and Beijing Is Freaking Out.' Protocol, March 19, 2021. https://www.protocol.com/china/chinese-deepfakes-regulators-alibaba-tencent.

Yeh, Chin-Yuan, Hsi-Wen Chen, Shang-Lun Tsai, and Shang-De Wang. 'Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks.' In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 53–62, 2020. https://doi.org/10.1109/WACVW50321.2020.9096939.

Yoon, Leonard, Dongseok Yang, Choongho Chung, and Sung-Hee Lee. 'A Mixed Reality Telepresence System for Dissimilar Spaces Using Full-Body Avatar.' In *SIGGRAPH Asia 2020 XR*, 1–2. SA '20. Virtual Event, Republic of Korea: Association for Computing Machinery, 2020. https://doi.org/10.1145/3415256.3421487.

Younus, Mohammed A., and Taha M. Hasan. 'Abbreviated View of Deepfake Videos Detection Techniques.' In *2020 6th International Engineering Conference 'Sustainable Technology and Development' (IEC)*, 115–20, 2020. https://doi.org/10.1109/IEC49899.2020.9122916.

Younus, Mohammed Akram, and Taha Mohammed Hasan. 'Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform.' In *2020 International Conference on Computer Science and Software Engineering (CSASE)*, 186–90, 2020. https://doi.org/10.1109/CSASE48920.2020.9142077.

Zakharov, Egor, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 'Few-Shot Adversarial Learning of Realistic Neural Talking Head Models.' *ArXiv:1905.08233 [Cs]*, September 25, 2019. http://arxiv.org/abs/1905.08233.

Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. 'StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks.' *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, no. 8 (August 2019): 1947–62. https://doi.org/10.1109/TPAMI.2018.2856256.

Zhang, Weiguo, Chenggang Zhao, and Yuxing Li. 'A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis.' *Entropy* 22, no. 2 (February 2020): 249. https://doi.org/10.3390/e22020249.

Zhang, Zhaohe, and Qingzhong Liu. 'Detect Video Forgery by Performing Transfer Learning on Deep Neural Network.' In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery*, edited by Yong Liu, Lipo Wang, Liang Zhao, and Zhengtao Yu, 415–22. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-32591-6_44.

Zhu, Bingquan, Hao Fang, Yanan Sui, and Luming Li. 'Deepfakes for Medical Video De-Identification: Privacy Protection and Diagnostic Information Preservation.' *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, February 7, 2020, 414–20. https://doi.org/10.1145/3375627.3375849.

QA-02-21-930-EN-N

The emergence of a new generation of digitally manipulated media – also known as deepfakes – has generated substantial concerns about possible misuse. In response to these concerns, this report assesses the technical, societal and regulatory aspects of deepfakes.

The rapid development and spread of deepfakes is taking place within the wider context of a changing media system. An assessment of the risks associated with deepfakes shows that they can be psychological, financial and societal in nature, and their impacts can range from the individual to the societal level. The report identifies five dimensions of the deepfake lifecycle that policy-makers could take into account to prevent and address the adverse impacts of deepfakes. The report includes policy options under each of the five dimensions, which could be incorporated into the AI legislative framework, the digital service act package and beyond. A combination of measures will likely be necessary to limit the risks of deepfakes, while harnessing their potential.

This is a publication of the Scientific Foresight Unit (STOA)
EPRS | European Parliamentary Research Service

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

ISBN 978-92-846-8400-7 | doi: 10.2861/325063| QA-02-21-930-EN-N