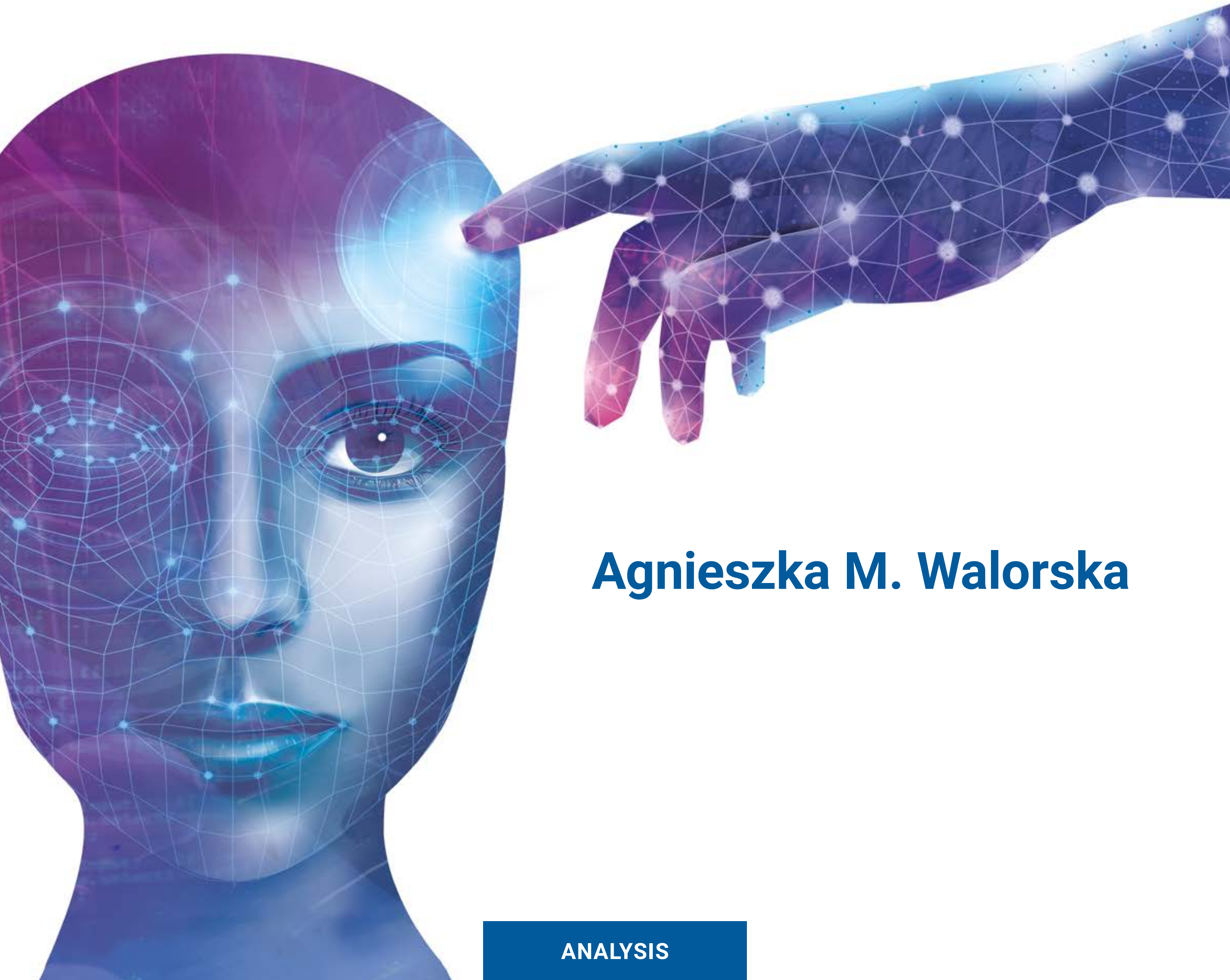
FRIEDRICH NAUMANN
FOUNDATION For Freedom.

DEEPFAKES
& DISINFORMATION



Agnieszka M. Walorska

ANALYSIS

IMPRINT

Publisher
Friedrich Naumann Foundation for Freedom
Karl-Marx-Straße 2
14482 Potsdam
Germany

/freiheit.org
/FriedrichNaumannStiftungFreiheit
/FNFreiheit

Author
Agnieszka M. Walorska

Editors
International Department
Global Themes Unit
Friedrich Naumann Foundation for Freedom

Concept and layout
TroNa GmbH

Contact
Phone: +49 (0)30 2201 2634
Fax: +49 (0)30 6908 8102
Email: service@freiheit.org

As of
May 2020

Photo Credits
Photomontages
© Unsplash.de, © freepik.de, P. 30 © AdobeStock

Screenshots
P. 16 © https://youtu.be/mSaIrz8lM1U
P. 18 © deepnude.to / Agnieszka M. Walorska
P. 19 © thispersondoesnotexist.com
P. 19 © linkedin.com
P. 19 © talktotransformer.com
P. 25 © gltr.io
P. 26 © twitter.com

All other photos
© Friedrich Naumann Foundation for Freedom (Germany)
P. 31 © Agnieszka M. Walorska

Notes on using this publication

This publication is an information service of the Friedrich Naumann Foundation for Freedom. The publication is available free of charge and not for sale. It may not be used by parties or election workers during the purpose of election campaigning (Bundestags-, regional and local elections and elections to the European Parliament).

Table of contents

EXECUTIVE SUMMARY

Applications of Artificial Intelligence (AI) are playing an increasing role in our society – but the new possibilities of this technology come hand in hand with new risks. One such risk is misuse of the technology to deliberately disseminate false information. Although politically motivated dissemination of disinformation is certainly not a new phenomenon, technological progress has made the creation and distribution of manipulated content much easier and more efficient than ever before. With the use of AI algorithms, videos can now be falsified quickly and relatively cheaply (*"deepfakes"*) without requiring any specialised knowledge.

The discourse on this topic has primarily focused on the potential use of deepfakes in election campaigns, but this type of video only makes up a small fraction of all such manipulations: in 96% of cases, deepfakes were used to create pornographic films featuring prominent women. Women from outside of the public sphere may also find themselves as the involuntary star of this kind of manipulated video (deepfake revenge pornography). Additionally, applications such as DeepNude allow static images to be converted into deceptively real nude images. Unsurprisingly, these applications only work with images of female bodies. But visual content is not the only type of content that can be manipulated or produced algorithmically. AI-generated voices have already been successfully used to conduct fraud, resulting in high financial damages, and GPT-2 can generate texts that invent arbitrary facts and citations.

What is the best way to tackle these challenges? Companies and research institutes have already invested heavily in technological solutions to identify AI-generated videos. The benefit of these investments is typically short-lived: deepfake developers respond to technological identification solutions with more sophisticated methods – a classical example of an arms race. For this reason, platforms that distribute manipulated content must be held more accountable. Facebook and Twitter have now self-imposed rules for handling manipulated content, but these rules are not uniform, and it is not desirable to leave it to private companies to define what *"freedom of expression"* entails.

The German federal government is clearly unprepared for the topic of *"Applications of AI-manipulated content for purposes of disinformation"*, as shown by the brief parliamentary inquiry submitted by the FDP parliamentary group in December 2019. There is no clearly defined responsibility within the government for the issue and no specific legislation. So far, only *"general and abstract rules"* have been applied. The replies given by the federal government do not suggest any concrete strategy nor any intentions of investing in order to be better equipped to deal with this issue. In general, the existing regulatory attempts at the German and European level do not appear sufficient to curb the problem of AI-based disinformation. But this does not necessarily have to be the case. Some US states have already passed laws against both non-consensual deepfake pornography and the use of this technology to influence voters.

Accordingly, legislators should create clear guidelines for digital platforms to handle deepfakes in particular, and disinformation in general, in a uniform manner. Measures can range from labelling manipulated content as such and limiting its distribution (excluding it from recommendation algorithms) to deleting it. Promoting media literacy should also be made a priority for all citizens, regardless of age. It is important to raise awareness of the existence of deepfakes among the general public and develop the ability of individuals to analyse audiovisual content – even though it is becoming increasingly difficult to identify fakes. In this regard, it is well worth taking note of the approach taken by the Nordic countries, especially Finland, whose population was found to be the most resilient to disinformation.

Still, there is one thing that we should not do: give in to the temptation of banning deepfakes completely. Like any technology, deepfakes do open up a wealth of interesting possibilities – including for education, film and satire – despite their risks.

GLOSSARY

Artificial General Intelligence / Strong AI

The concept of strong AI or AGI refers to a computer system that masters a wide range of different tasks and thereby achieves a human-like level of intelligence. Currently, no such AI application exists. For instance, no single system is currently able to recognise cancer, play chess and drive a car, even though there are specialised systems that can perform each task separately. Multiple research institutes and companies are currently working on strong AI, but there is no consensus on whether it can be achieved, and, if so, when.

Big Tech

The term *"Big Tech"* is used in the media to collectively refer to a group of dominant companies in the IT industry. It is often used interchangeably with *"GAFA"* or *"the Big Four"* for Google, Apple, Facebook, and Amazon (or "GAFAM" if Microsoft is included). For the Chinese big tech companies, the abbreviation BATX is used, for Baidu, Alibaba, Tencent, and Xiaomi.

Cheapfakes / Shallowfakes

In contrast to deepfakes, shallowfakes are image, audio or video manipulations created with relatively simple technologies. Examples include reducing the speed of an audio recording or displaying content in a modified context.

DARPA

The Defense Advanced Research Projects Agency is part of the US Department of Defense, entrusted with the task of researching and funding groundbreaking military technologies. In the past, projects funded by DARPA have resulted in major technologies that are also used in non-military applications, including the internet, machine translation and self-driving vehicles.

Deepfake

Deepfakes (a portmanteau of deep learning and fake) are the product of two AI algorithms working together in a so-called Generative Adversarial Network (GAN). GANs are best described as a way to algorithmically generate new types of data from existing datasets. For example, a GAN could analyse thousands of pictures of Donald Trump and then generate a new picture that is similar to the analysed images but not an exact copy of any of them. This technology can be applied to various types of content – images, moving images, sound, and text. The term deepfake is primarily used for audio and video content.

Deep Learning

Deep learning is a sub-area of machine learning, where artificial neural networks learn from large amounts of data. Similar to humans learning from experience, deep learning algorithms repeat a task to gradually improve the results. This is called deep learning because the neural networks have multiple layers to enable learning. Deep learning allows machines to solve complex problems, even when using non-uniform, unstructured datasets.

Deep Porn

Deep porn refers to the use of deep learning methods to generate artificial pornographic images.

Generative Adversarial Network

Generative adversarial networks are algorithmic architectures based on a pair of two neural networks, namely one generative network and one discriminatory network. The two networks compete against one another (the generative network generates data and the discriminatory network falsifies the data) to generate new synthetic datasets. The process is repeated multiple times to achieve results that are extremely similar to real data. The networks can work with different types of data and can therefore be used for generating images, as well as text, audio or video.

GPT-2

GPT-2 is a framework based on an artificial neural network developed by the research company OpenAI. It is able to automatically generate English-language texts. The dataset on which GPT-2 is based contains around 45 million pages of text. Unlike conventional text generators, GPT-2 does not compose texts from finished text blocks, and it is not restricted to any specific domain. It can generate new content from any given sentence or section of text.

GLOSSARY

IBM Watson

IBM Watson is a system based on machine learning developed by IBM. It was developed with the goal of creating a system that can understand and answer questions asked in natural language. Watson received widespread media attention in 2011 when it beat the best human players on the television quiz show Jeopardy. Since then, IBM Watson has been marketed as *"AI for business"*, offering a range of cloud and data products for various industries – from healthcare to film production.

AI Winter

An AI winter is a period of declining interest and decreasing research funding in the field of artificial intelligence. The term was coined by analogy with the idea of nuclear winter. As a technological field, AI has experienced several phases of hype since the 1950s, followed by disappointment, criticism and funding cuts.

Artificial Neural Networks

Artificial Neural Networks (ANNs) are computer systems loosely inspired by the biological neural networks found in the brains of humans and animals. ANNs *"learn"* how to perform tasks based on examples without being programmed with any task-specific rules.

ANNs can, for example, learn to identify images containing cats by analysing sample images that have manually been labelled as *"cat"* or *"no cat"* and then use the results to identify cats in other images.

Machine Learning

Fundamentally, machine learning is a method that applies algorithms to analyse data, to learn from this data and then make predictions based on it. Thus, rather than manually programming software with precisely defined instructions to perform a certain task, the software is trained with large amounts of data and algorithms that give it the ability to learn how the task should be performed.

Microtargeting

Microtargeting is a digital marketing technique that seeks to match ad campaigns with the consumers or businesses most likely to be interested in the product or service. Depending on the platform, this is done using, for example, target audience demographics, interests and browsing history. Based on these criteria, different recipients can be addressed in completely different ways by the same campaign organiser. This marketing tool was originally developed for political campaigns but is today also used in commercial campaigns.

Phishing

Phishing is a cyberattack method using email. The goal is to convince the email recipient that the message is authentic and relevant (e.g. a notification from their bank) to motivate them to click on a link or download an attachment. This allows hackers to gain access to sensitive information such as passwords.

Revenge Porn

Revenge pornography refers to the sharing of intimate sexual images or videos without the consent of the participants. This is frequently done as a form of revenge by ex-partners after the end of a relationship. Three quarters of the victims of revenge pornography are women.

Weak AI or Specialised AI

The algorithms of weak AI are specialised in performing very specific tasks, such as recognising faces, understanding language or playing chess. Although they are typically much better or more efficient than humans at these tasks, they are only capable of completing the problems for which they were designed. Every modern application of artificial intelligence belongs to the category of weak AI, even complex-seeming systems such as self-driving vehicles or language assistants.

Social Engineering

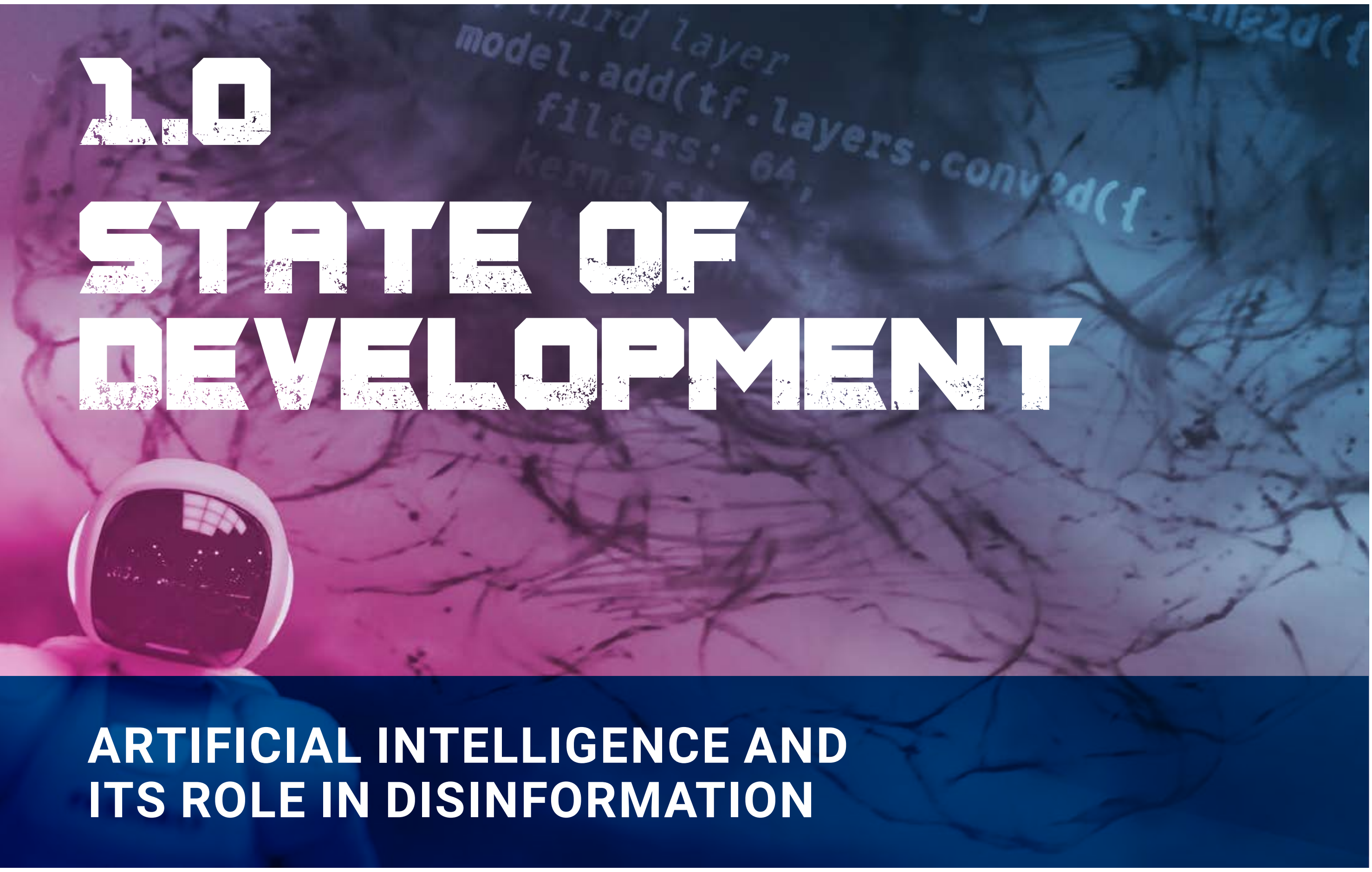
Social engineering refers to any measures involving the targeted influencing of a person or people, for example to gain access to confidential information or to convince a target to make a payment. This practice is also called *"social hacking"* when the goal is to gain access to the computer systems of the target person or organisation.

Superintelligence

Superintelligence is a hypothetical scenario in which artificial intelligence not only surpasses the most intelligent people as individuals but supersedes the collective intelligence of humankind.

Vishing

Vishing (voice phishing) is a phishing method that uses telephone calls instead of email. Using deepfakes to generate voice clips can improve the effectiveness of this technique.



1.0 STATE OF DEVELOPMENT

ARTIFICIAL INTELLIGENCE AND ITS ROLE IN DISINFORMATION

Although the roots of the technology stretch back to the mid-20th century, artificial intelligence received little attention for a long time. The long AI winter only began to abate in the early 2010s. In 2011, IBM's computer system Watson beat the best human players in the television quiz show Jeopardy[1], Google's self-driving car prototypes travelled more than 100,000 miles (160,000 kilometres) and Apple introduced their *"smart personal assistant"* Siri. Since then, public interest in artificial intelligence, and especially in the risks associated with it, has been steadily growing. The discourse on superintelligence – triggered by a book of the same title by Nick Bostrom published in 2014 – generated even more attention. Prominent personalities have since repeatedly warned about AI, sometimes taking on an alarming tone. Stephen Hawking (*"The development of full artificial intelligence could spell the end of the human race."*) and Elon Musk *("AI is a fundamental existential risk for human civilisation.")* are frequently cited. While superintelligence and so-called *"strong AI"* (AGI, Artificial General Intelligence) are still in the distant future, *"weak AI"* and its arguably not-so-weak algorithms are already playing a steadily expanding role in business, society and politics. The author's opinion is that the effects on health, energy, security, mobility and many other areas will be largely positive. However, we will only be able to enjoy the positive aspects of these developments if we recognise the risks associated with this technology and successfully counteract them.

One such risk is misuse of the technology to deliberately disseminate false information. Of course, politically motivated disinformation is not a new phenomenon. Stalin and Mao are the most prominent examples of dictators who regularly ordered their photographs

[1] Jeopardy is a quiz game show where participants receive general clues presented as an answer and must formulate their response in the form of a question. Over the years, German adaptations have included *"Riskant on RTL"* and *"Der Große Preis on ZDF"*.

> ***"However, we will only be able to enjoy the positive aspects of these developments if we recognise the risks associated with this technology and successfully counteract them."***

to be edited to ensure that old images would be consistent with the latest *"truth"*: anyone who had fallen out of favour was removed from pictures, new additions to the party leadership were retroactively edited in; even the context of pictures was modified, for example by changing the background. The goal of manipulating these visual records was to create new facts, to rewrite past events and history itself.

Historically, performing these modifications was tedious and required specialised knowledge; today, with the right smartphone app, anybody can do the same effortlessly. And the technology has not stopped at photography. Producing a fake video that appears believable still requires a fair deal of effort. But certain methods of artificial intelligence are making it increasingly easy to manipulate existing videos. These videos have become known as *"deepfakes"*. They are still relatively uncommon on the internet, but as their use and dissemination increases, they are turning into a growing challenge for our society. Not only does manipulated content spread very quickly on platforms such as Facebook or YouTube, it is also specifically targeted towards users who are receptive to it. Furthermore, the spread of disinformation is increasingly shifting towards messenger services such as WhatsApp. There, encrypted messages are distributed over private connections, this increases the trust in the forwarded information, creating a kind of hidden virality. Encryption of private online communications is a desirable commodity, similar to the secrecy of written letters – it prevents messages from being viewed by third parties. But encryption also means that any disseminated information cannot be checked for truthfulness and moderated accordingly.

2.0 CHEAPFAKES & DEEPFAKES

TECHNOLOGICAL POSSIBILITIES FOR THE MANIPULATION OF TEXT, IMAGES, AUDIO AND VIDEO

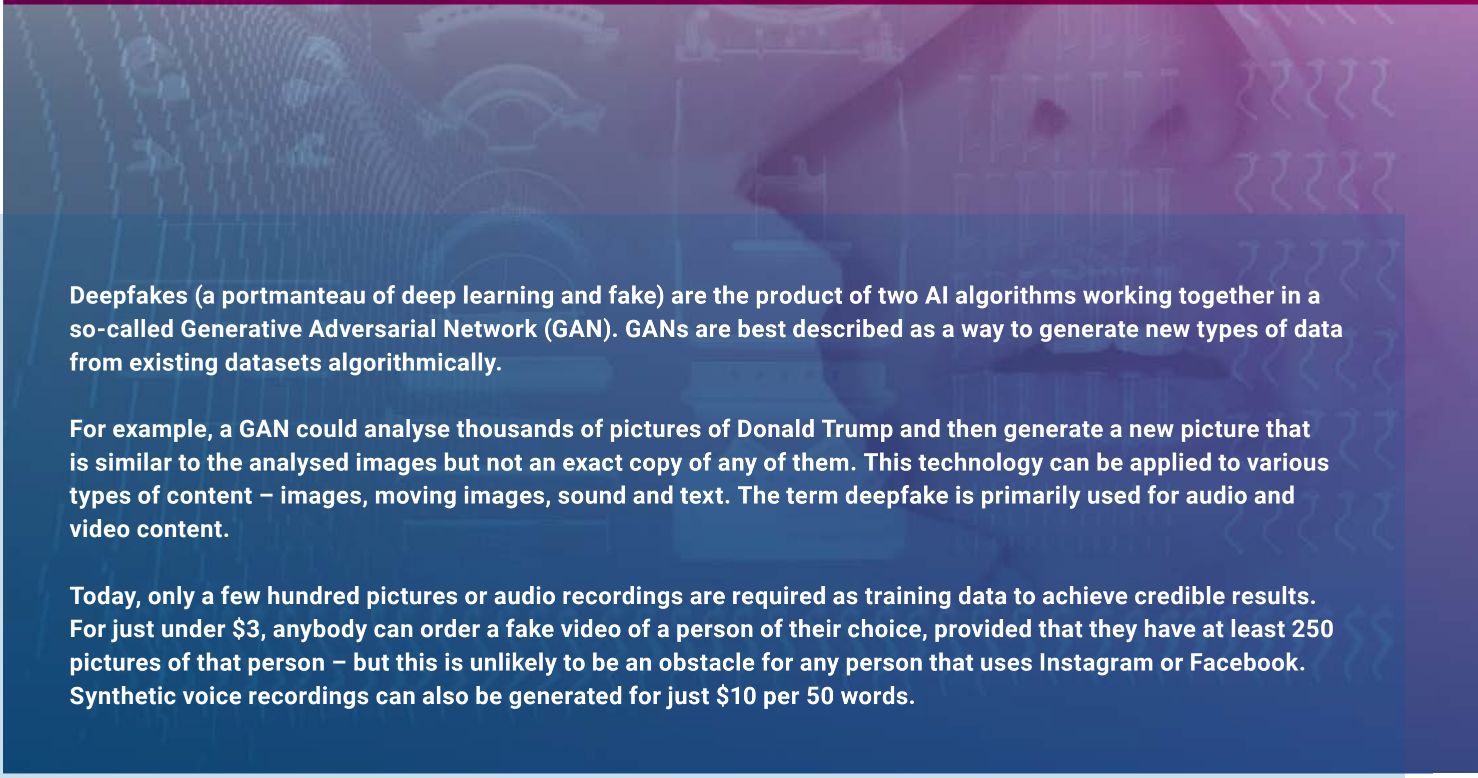
Over the past two years, the term deepfake has become increasingly widespread. But what exactly are deepfakes, and how are they different from other manipulated content? Although the first scientific AI-based experiments on video manipulation go back to the late 1990s, the general public only became aware of the technical possibilities towards the end of 2017.

This was also when the terminology was coined, when a Reddit user named *"Deepfakes"* and other members of the Reddit community *"r/deepfakes*" published content created by them.

Unsurprisingly, in many cases, this has been used to make pornographic videos where the faces of the actresses are replaced by celebrities such as Scarlett Johansson or Taylor Swift. A more harmless example involved taking film scenes and replacing the face of each actor with Nicolas Cage.

> ***"Unsurprisingly, in many cases, this has been used to make pornographic videos where the faces of the actresses are replaced by celebrities such as Scarlett Johansson or Taylor Swift."***

DEEPFAKES WORK AS FOLLOWS



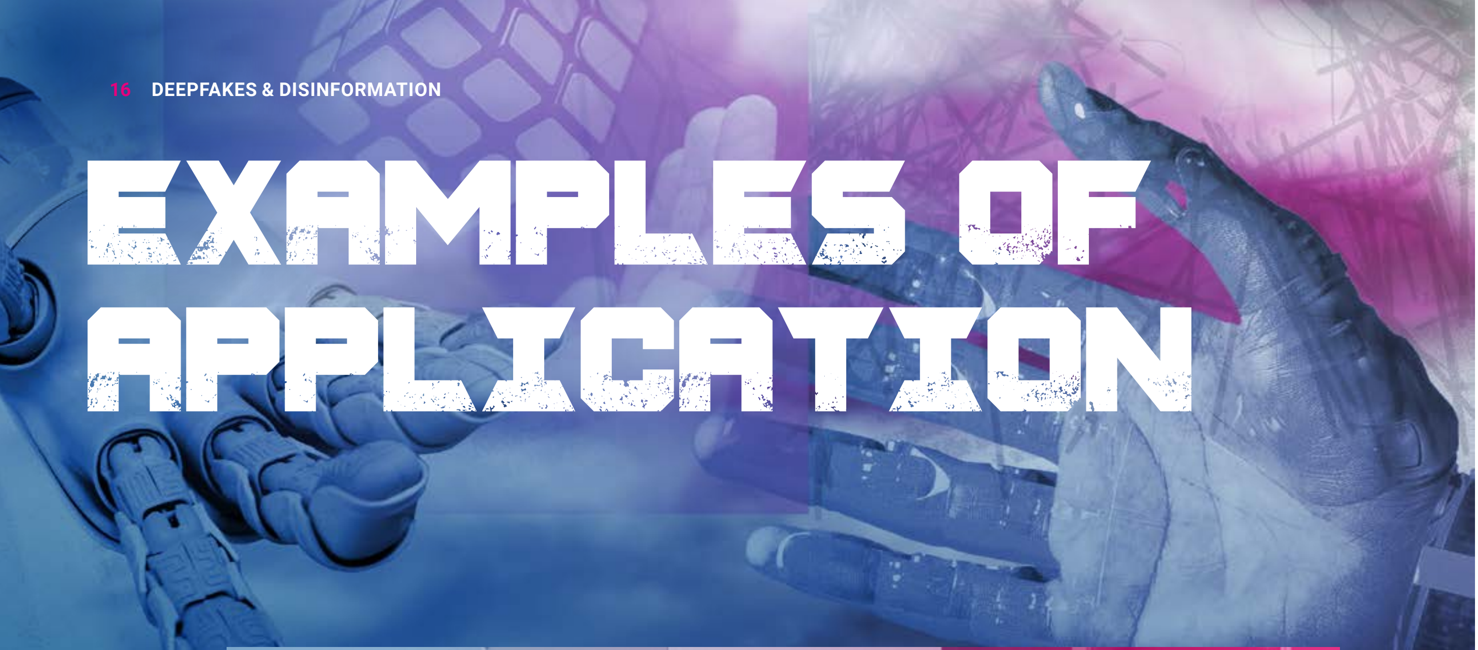
Deepfakes (a portmanteau of deep learning and fake) are the product of two AI algorithms working together in a so-called Generative Adversarial Network (GAN). GANs are best described as a way to generate new types of data from existing datasets algorithmically.

For example, a GAN could analyse thousands of pictures of Donald Trump and then generate a new picture that is similar to the analysed images but not an exact copy of any of them. This technology can be applied to various types of content – images, moving images, sound and text. The term deepfake is primarily used for audio and video content.

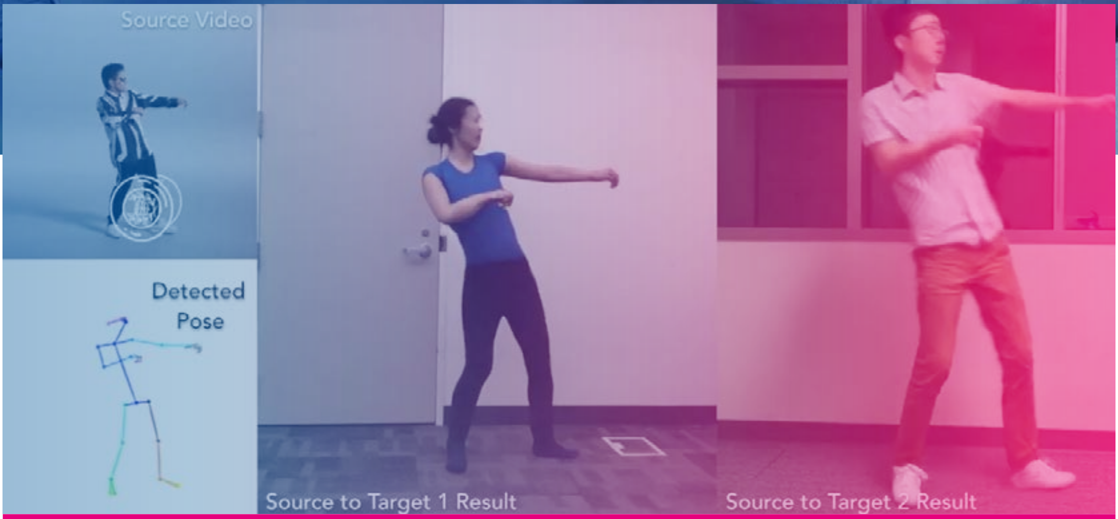
Today, only a few hundred pictures or audio recordings are required as training data to achieve credible results. For just under $3, anybody can order a fake video of a person of their choice, provided that they have at least 250 pictures of that person – but this is unlikely to be an obstacle for any person that uses Instagram or Facebook. Synthetic voice recordings can also be generated for just $10 per 50 words.

2.1 Deepfakes vs Cheapfakes

Although pornographic manipulations are undoubtedly the most common examples of deepfakes, they are not the primary motivation for the current societal debate. Interestingly, the video that sparked the debate was not a deepfake by any means, but simply a cheapfake (sometimes also called a shallowfake): a video of the speaker of the US House of Representatives, Nancy Pelosi, faked with very simple technical means. The recording was slowed to around 75% of its original speed, while raising the pitch so the voice still sounded natural. The results: The viewer was given a plausible impression that Nancy Pelosi was drunk. The video was shared millions of times on social media. This shows how even the simplest forgeries can distort reality and be exploited for political purposes. Nevertheless, it was historically very difficult to falsify recordings to make the subject perform completely different movements or speak completely different words than in the original video. Until now.



EXAMPLES OF APPLICATION



Source: https://youtu.be/mSaIrz8lM1U

2.2 Manipulation of movement patterns

In 2018, an application by four Berkeley researchers attracted widespread attention, using artificial intelligence to transfer the dance routine of a source person (such as a professional dancer) to a target person.[2)]

The movements are transferred from the source video to a *"stick figure"*. The neural network then synthesizes the target video according to the *"stick figure movements"*.

The result is a *"faked"* video where a third person dances like a professional. Of course, this type of algorithm could be used not only to imitate dance movements, but potentially to generate any other form of movement. This opens the door to portraying political opponents in compromising situations: What would, for instance, be the ramifications of a video showing a politician performing a Nazi salute or even just giving the middle finger?

2) https://arxiv.org/pdf/1808.07371.pdf

ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) are computer systems loosely inspired by the biological neural networks found in the brains of humans and animals.

ANNs *"learn"* how to perform tasks based on examples without being programmed with any task-specific rules. They can, for example, learn to identify images containing cats by analysing sample images that have manually been labelled as *"cat"* or *"no cat"* and use the results to identify cats in other images.

Voice and facial expressions

Forgeries can have even further-reaching consequences by making individuals appear to speak words that were never said, accompanied by gestures, facial expressions and voice impressions that seem incredibly realistic. A series of such videos were created, including examples of Barack Obama and Mark Zuckerberg, not to deceive the audience, but to demonstrate the possibilities and risks of this technology. Since then, there has been an instance where a deepfake was created and distributed by a political party, the Belgian Socialistische Partij Anders (sp.a.).

In May 2018, the party posted a video on Facebook in which Trump mocked Belgium for observing the Paris climate agreement.[3] Despite obviously poor quality and unnatural mouth movements that should rouse the suspicion of any attentive viewer, the video triggered hundreds of comments, many of them expressing outrage that the American president would dare to meddle in Belgian climate policy. The creators of this video were also trying to promote understanding of an issue. The video was a targeted provocation to draw people's attention to an online petition calling for the Belgian government to take more urgent action on climate issues. But what if someone created a video where Trump talks about a topic other than Belgian climate policy, for example his intent to attack Iran?

[3] https://www.facebook.com/watch/?v=10155618434657151

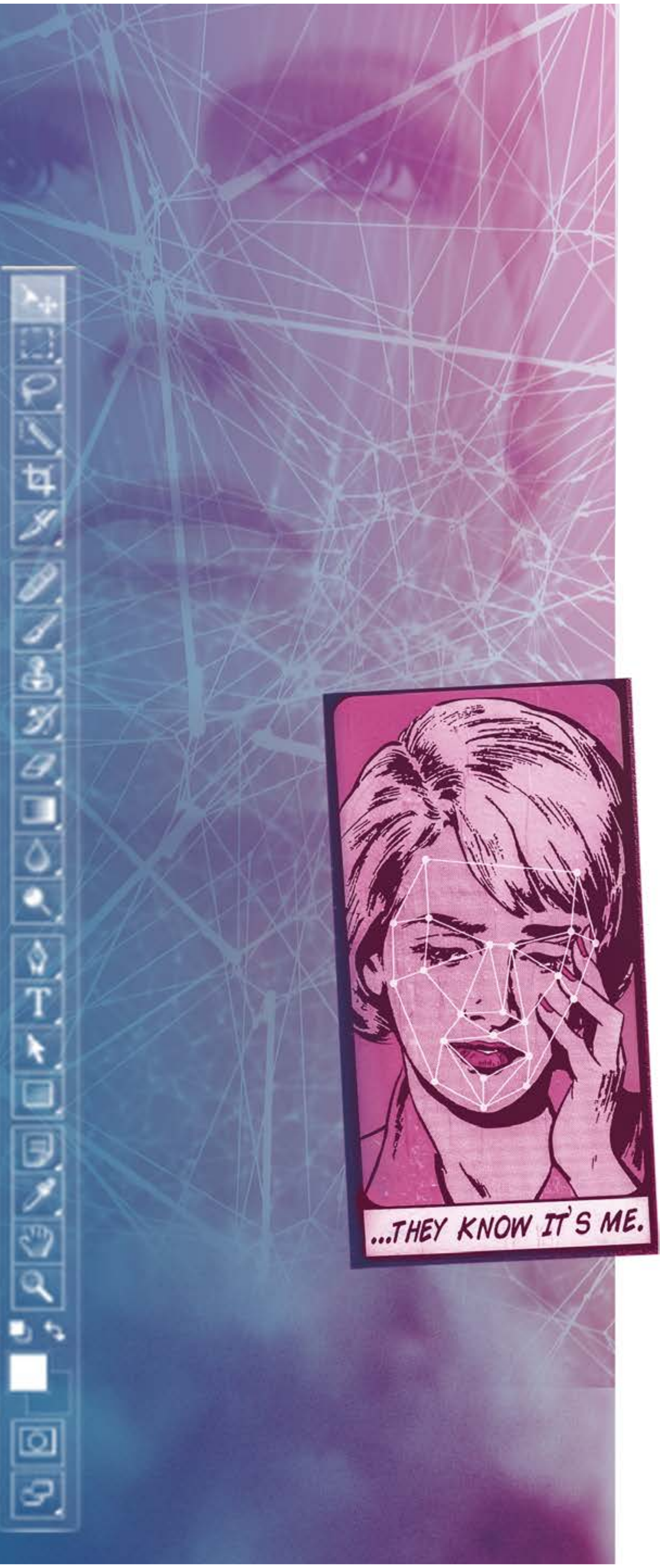


Image manipulation: DeepNude and artificial faces

Image and text content are often not categorised as deepfakes, although they can be generated with very similar technology. There is a simple reason for this: both images and texts can be manipulated so easily without requiring complex technology that the *"benefit"* (or harm, depending on the perspective) of doing so is much smaller than for manipulations of audio and video content. Furthermore, video recordings are much more effective than text and static images at triggering emotions such as fear, anger or hate.

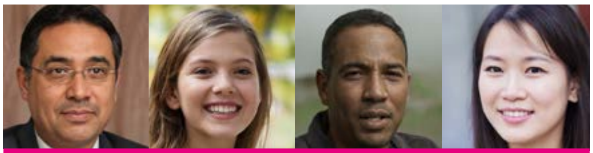
Nevertheless, some examples of AI-based manipulated picture/text content have also attracted attention. As for videos, the primary purpose of image manipulation algorithms is to create fake pornographic content. Applications like DeepNude can convert a bikini photo into a very realistic nude image in a matter of seconds.

Unsurprisingly, the app only works with women (any attempt to select a male image simply generates female genitalia). But this makes each and every woman a potential victim of *"revenge porn"*, even if no real naked pictures ever existed.

A nude picture generated with the deepnude.to application

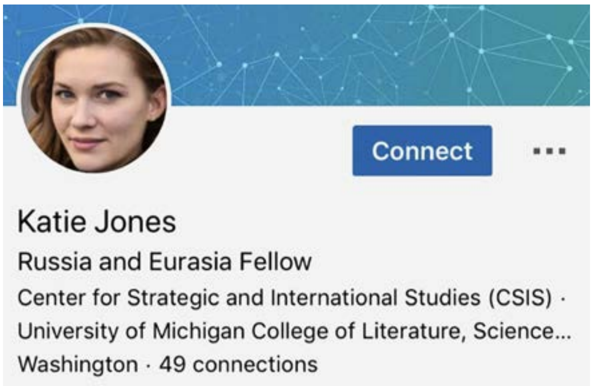
These neural networks are not restricted to the manipulation of images of real people. They can also *"create"* completely new people – or at least completely new faces.

The commercial applications of this technology are obvious: image databases can be populated more cost-efficiently using AI rather than real people. But this also means that creating fake social media profiles, for example with the purpose of spreading political content, becomes significantly easier.



Faces randomly generated by thispersondoesnotexist.com

There have also been suspected attempts of espionage with computer-generated profile pictures, for example the LinkedIn profile of one *"Katie Jones"*, an alleged researcher working at a US think tank.



Katie Jones
Russia and Eurasia Fellow
Center for Strategic and International Studies (CSIS) · University of Michigan College of Literature, Science...
Washington · 49 connections

LinkedIn profile: AI-generated profile for *„Katie Jones"*

Before expert analysis identified several visual anomalies suggesting that the image was synthetic, the profile successfully connected with 52 political figures in Washington, including a deputy assistant secretary of state, a senior adviser to a senator and a prominent economist.[4]

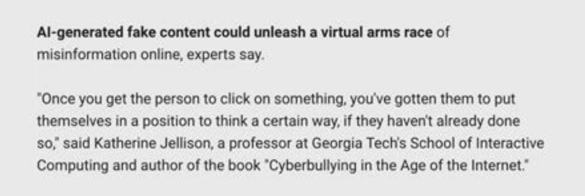
The account was quickly removed by LinkedIn but is thought to have belonged to a network of phantom profiles, some of which may still exist, that could be used for phishing attacks.

AI-generated texts

The application described above can be implemented particularly effectively in combination with AI-driven text generation.

Many people may already have heard of this possibility thanks to the GPT-2 text generator created by the research company OpenAI. Due to the potential for abuse, GPT-2 was originally considered too dangerous to be made available to the general public.[5] The company later decided to publish GPT-2 in several stages, since its creators have so far been unable to find any clear evidence of misuse.[6]

Even if there has not yet been misuse, the creators acknowledge that people would largely find the text generated by GPT-2 credible, that the generator could be fine-tuned to produce extremist content, and that identifying generated text would be challenging. With the *"Talk To Transformer"* application, anybody can try out GPT-2 for themselves.



AI-generated fake content could unleash a virtual arms race of misinformation online, experts say.

"Once you get the person to click on something, you've gotten them to put themselves in a position to think a certain way, if they haven't already done so," said Katherine Jellison, a professor at Georgia Tech's School of Interactive Computing and author of the book "Cyberbullying in the Age of the Internet."

Highlighted text – input, remaining text – generated by AI with www.talktotransformer.de

Entering one or more sentences into the generator outputs a piece of text beginning with the submitted input. The results are often – but not always – surprisingly coherent. They strike the same tone as the input and simulate credibility by inventing experts, statistics and quotes.

[4] https://www.cnet.com/news/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/

[5] https://openai.com/blog/better-language-models/

[6] https://openai.com/blog/gpt-2-1-5b-release/

3.0 DISSEMINATION & CONSEQUENCES

HOW DANGEROUS ARE DEEPFAKES IN REALITY?

3.1 Dissemination

It is difficult to precisely quantify the dissemination of deepfakes, especially since their number is undoubtedly steadily growing.

Deeptrace, a company that offers a technological solution to detect deepfakes, attempted to give a precise estimate in their report: The State of Deepfakes: Landscape, Threats, and Impact.[7)]

Published in September 2019, the report estimates that the number of deepfakes almost doubled in seven months from 7,964 in December 2018 to 14,678 in July 2019. Of these deepfakes, 96% were non-consensual pornographic content that exclusively depicted the female body.

The primary victims were prominent women, for whom thousands of fake pictures can be found online. According to the Deeptrace report, the four most popular deep porn websites alone registered more than 134 million views of fake videos of female celebrities. But many private individuals are also affected by the phenomenon of revenge pornography mentioned above. The increase is driven primarily by greater accessibility to tools and services that allow deepfakes to be created without any knowledge of programming.

In 2019, there were also reports of AI-generated language clones being used for social engineering. In August, The Wall Street Journal reported [8)] on the first case of AI-based voice fraud – also known as vishing (short for *"voice phishing"*) – at a cost of €220,000 for the German company that was targeted.

The software imitated the voice of the German manager so successfully, including his intonations and slight German accent, that his British colleague immediately complied with the caller's urgent request to transfer the stated amount. Although this is currently an isolated incident, it seems likely that there will be more such attempts in the future.

A significant part of the media coverage of deepfakes has focused on their potential to discredit political opponents and undermine democratic processes. So far, this potential has not materialised. Although there have been technically manipulated videos of politicians such as Barack Obama, Donald Trump and Matteo Renzi, they were motivated primarily by satire or created for demonstration purposes, and their falseness was quickly disclosed.

„Although there have been technically manipulated videos of politicians such as Barack Obama, Donald Trump and Matteo Renzi, they were primarily motivated by satire or created for demonstration purposes, and their falseness was quickly disclosed.“

3.2 Consequences

However, the fact that politicians have not yet used deepfakes for disinformation does not mean that deepfakes have not already influenced the political discourse. One example that received little attention in the Western media demonstrates how the simple knowledge of the existence of deepfakes can affect the political climate.

The president of Gabon, Ali Bongo, did not appear in public for months after experiencing a stroke. Unsurprisingly, rumours began spreading that the president had passed away. To quash the speculation, the president published a video in December 2018 to give his usual New Year's speech. But the recording had the opposite effect. Many people thought that Bongo looked strange and immediately suspected that the video was fake. Shortly afterwards, the military launched a failed coup, citing the supposed deepfake as one of their motives.[9)]

However, subsequent forensic analysis confirmed that the recording was authentic. Ali Bongo has since recovered from his stroke and remains in office.

This shows that the biggest threat posed by deepfakes isn't the deepfakes themselves. The mere fact that such videos are technically possible raises the question: Can we still trust the authenticity of videos?

This question will cast a shadow over the 2020 US presidential elections. In the 2016 election campaign, AI-supported disinformation and manipulation, most prominently in the form of microtargeting and bots, had already begun to play a role. Deepfakes now represent another instrument in the arsenal of disinformation. Even if few or no deepfakes are actually used in the election campaign, it is likely that many politicians will gratefully accept the opportunity to shrug off real but unfavourable recordings as forgeries.

[7)] The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019.

[8)] https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

[9)] https://www.technologyreview.com/s/614526/the-biggest-threat-of-Deepfakes-isnt-the-Deepfakes-themselves/

3.3 Are there any examples of positive applications of deepfakes?

"Technology is continually giving us ways to do harm and to do well; it's amplifying both. [...] But the fact that we also have a new choice every time is a new good," [10] says Kevin Kelly, the long-standing editor-in-chief and founding member of the technology magazine Wired. Might this statement also apply to deepfakes?

The technology is especially promising for the film industry, particularly in post-production and dubbing. Why? Currently, modifying a piece of dialogue retroactively is very expensive for film studios. The actors, film crew and film set need to be rebooked. The technology behind deepfakes could allow these types of changes to be made quickly and at a fraction of the cost.

Significant improvements could also be made to film dubbing. It would become possible to adapt the lip movements of the actors to the dubbed words or synthesise their voices to adapt them to the target language, meaning that dubbing is no longer necessary.

One example of such an application is a video by David Beckham promoting a campaign against malaria.[11] He *"speaks"* in several languages – and his mouth appears to synchronise perfectly with the words in each case.

[10] **Quote from** https://www.edge.org/conversation/kevin_kelly-the-technium/

[11] https://www.malariamustdie.com/

Education is another interesting area of application: videos of historical figures could, for example, be created to tell their story or answer questions. The project *"Dimensions of History"* [12] by the Shoah Foundation of the University of Southern California attracted a lot of media attention, featuring interviews and holographic recordings of 15 holocaust survivors. This travelling exhibition was displayed in various museums throughout the US and was most recently hosted by the Swedish Museum of History.

Visitors to the exhibition were given the opportunity to ask the holograms questions. The speech recognition software then matched their question with a segment of the interview. With deepfake technology, this could be implemented on a larger scale, in multiple languages.

„The technology is especially promising for the film industry, particularly in post-production and dubbing.“

[12] https://sfi.usc.edu/dit

4.0 FACING DEEPFAKES

HOW CAN WE FACE THE CHALLENGES ASSOCIATED WITH DEEPFAKES?

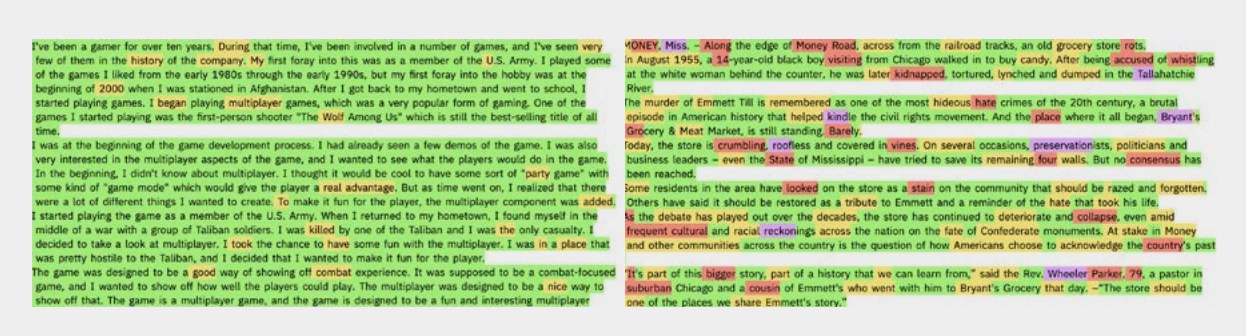
These positive examples are of course not intended to minimise the potential dangers posed by deepfakes. The risks are undisputed and require decisive countermeasures – on this, there is a consensus. But there is less agreement on the exact nature of these countermeasures. Also, the question arises of how to guarantee the rights of individuals to freedom of expression without undermining society's need for a reliable information system.

4.1 Technological solutions for identifying and combating deepfakes

One approach to combating counterfeiting is to develop technologies that are capable of distinguishing between fake content and real content. This approach uses algorithms similar to those which generated the fakes in the first place. Using GLTR, a model based on the GPT-2 system mentioned above, researchers from the MIT-IBM Watson AI Lab and HarvardNLP investigated whether the same technology used to write independently fabricated articles can be used to recognise text passages that were generated by AI. When a text passage is generated in the test application, its words are highlighted in green, yellow, red or purple to indicate their predictability, in decreasing order. The higher the proportion of words with low predictability, namely sections marked in red and purple, the greater the likelihood that the passage was written by a human author. The more predictable the words (and the *"greener"* the text), the more likely the text was automatically generated.

Similar techniques could be used to expose manipulated videos. In 2018, researchers observed that the actors in deepfake videos didn't blink. This was because the static images used to generate the videos primarily showed people whose eyes were open. But the usefulness of this observation was short-lived. As soon as this information became public, videos began to appear with blinking people. A similar trend can be expected for any other identification mechanisms discovered in the future. This game of cat-and-mouse has been underway in the cybersecurity field for decades – progress always benefits both sides.

But this doesn't mean that efforts to identify deepfakes should be discontinued. In September 2019, Facebook – in collaboration with the PAI initiative[13], Microsoft and several universities – announced a *"Deepfake Detection Challenge"*[14] endowed with a $10 million prize.



I've been a gamer for over ten years. During that time, I've been involved in a number of games, and I've seen very few of them in the history of the company. My first foray into this was as a member of the U.S. Army. I played some of the games I liked from the early 1980s through the early 1990s, but my first foray into the hobby was at the beginning of 2000 when I was stationed in Afghanistan. After I got back to my hometown and went to school, I started playing games. I began playing multiplayer games, which was a very popular form of gaming. One of the games I started playing was the first-person shooter "The Wolf Among Us" which is still the best-selling title of all time.

I was at the beginning of the game development process. I had already seen a few demos of the game. I was also very interested in the multiplayer aspects of the game, and I wanted to see what the players would do in the game. In the beginning, I didn't know about multiplayer. I thought it would be cool to have some sort of "party game" with some kind of "game mode" which would give the player a real advantage. But as time went on, I realized that there were a lot of different things I wanted to create. To make it fun for the player, the multiplayer component was added.

I started playing the game as a member of the U.S. Army. When I returned to my hometown, I found myself in the middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.

The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer

MONEY, Miss. – Along the edge of Money Road, across from the railroad tracks, an old grocery store rots.

In August 1955, a 14-year-old black boy visiting from Chicago walked in to buy candy. After being accused of whistling at the white woman behind the counter, he was later kidnapped, tortured, lynched and dumped in the Tallahatchie River.

The murder of Emmett Till is remembered as one of the most hideous hate crimes of the 20th century, a brutal episode in American history that helped kindle the civil rights movement. And the place where it all began, Bryant's Grocery & Meat Market, is still standing. Barely.

Today, the store is crumbling, roofless and covered in vines. On several occasions, preservationists, politicians and business leaders – even the State of Mississippi – have tried to save its remaining four walls. But no consensus has been reached.

Some residents in the area have looked on the store as a stain on the community that should be razed and forgotten. Others have said it should be restored as a tribute to Emmett and a reminder of the hate that took his life.

As the debate has played out over the decades, the store has continued to deteriorate and collapse, even amid frequent cultural and racial reckonings across the nation on the fate of Confederate monuments. At stake in Money and other communities across the country is the question of how Americans choose to acknowledge the country's past

"It's part of this bigger story, part of a history that we can learn from," said the Rev. Wheeler Parker, 79, a pastor in suburban Chicago and a cousin of Emmett's who went with him to Bryant's Grocery that day. –"The store should be one of the places we share Emmett's story."

Results of the analysis: human author vs. text generator, Source: gltr.io

Facebook also commissioned a dataset with images and videos by actors specifically recorded for this purpose, so that the challenge would have adequate data to work with. A few weeks later, Google also released a dataset containing 3,000 manipulated videos with the same goal.

The US research funding agency DARPA has also been working on recognising manipulated content as part of the MediFor programme (short for Media Forensics) since 2016, investing more than $68 million over two years.[15] Little information is available on whether – and if so what type of – technical solutions to combat deepfakes are being developed in Germany and Europe.

Most measures are being undertaken by individual companies, such as Deeptrace mentioned above, as well as research projects like Face2Face by Matthias Nießner[16], a professor at the Technical University of Munich.

According to the response of the German government to a parliamentary question submitted by the FDP parliamentary group, the "National Research Centre for Applied Cybersecurity" CRISP/ATHENE is currently working on this issue with the Technical University of Munich and the Fraunhofer Institute.

In addition, the German international broadcaster Deutsche Welle (DW), the Fraunhofer Institute for Digital Media Technology (IDMT) and the Athens Technology Centre (ATC) have initiated the joint research project *"Digger"*. The goal of this project is to expand the web-based verification platform *"Truly Media"* by DW and the ATC with audio forensic technology by the Fraunhofer IDMT, among other things, to offer assistance to journalists.[17] However, this response does not suggest any concrete strategy nor intentions of investing in this topic by the federal government.

> ***"The higher the proportion of words with low predictability, namely sections marked in red and purple, the greater the likelihood that the passage was written by a human author. The more predictable the words (and the 'greener' the text), the more likely the text was automatically generated."***

[13] The Partnership on AI (PAI) is an organisation uniting universities, researchers, NGOs and enterprises to gain a better understanding of the impacts of AI and its effects on society. www.partnershiponai.org

[14] https://ai.facebook.com/blog/deepfake-detection-challenge/

[15] https://futurism.com/darpa-68-million-technology-Deepfakes

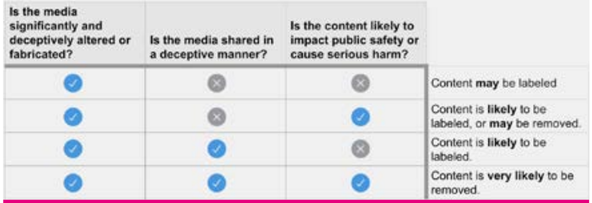
[16] https://niessnerlab.org/projects/thies2016face.html

[17] https://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf

4.2 Self-regulation attempts by social media platforms

Although big tech companies have contributed data and financial resources towards a technological solution to this problem, calls for Facebook and similar companies to take additional measures have been intensifying, since their platforms are key in the spread of disinformation. In response, Twitter and Facebook released statements about their plans to address deepfakes in late 2019 and early 2020, respectively.

In November 2019, Twitter asked its users for feedback on a *"policy proposal for synthetic and manipulated media"*. Guidelines were then announced at the beginning of February 2020: any photo, audio or video that has been *"significantly altered or falsified"* with the goal of misleading people would be removed if Twitter believes that it may cause serious harm – for example by endangering the physical security of individuals or prompting *"widespread civil unrest"*. If not, the tweets may still be labelled as manipulated media, showing a warning when the content is shared, and deprioritising the content in user feeds. These changes are to take effect on 5 March, 2020.[18]



Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Twitter: new approach to synthetic and manipulated media: https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

> ***“Little information is available on whether – and if so what type of – technical solutions to combat deepfakes are being developed in Germany and Europe.“***

Facebook is going one step further. On 6 January, 2020, Monika Bickert, Facebook’s Vice President of Global Policy Management, announced in a blog post that deepfakes meeting certain criteria would henceforth be deleted from the platform.[19] According to the blog post, any content modified or synthesised using AI in such a way that it appears authentic to the average person would be deleted. However, satirical content is excluded from these guidelines, which leaves significant room for interpretation.

Interestingly, the guidelines do not apply to cheapfakes; they explicitly and exclusively target AI-generated content. Accordingly, the fake video of Nancy Pelosi mentioned earlier continues to be available on Facebook.[20] Although Facebook admitted that its fact-checkers had flagged the video as fake, it declined to delete it because the company *“does not enforce a policy that requires information posted on Facebook to be truthful”*.[21]

This approach reflects Facebook’s position on freedom of expression and goes beyond the issue of deepfakes. In the debate on political advertising, Rob Leathern, the Director of Product Management at Facebook, wrote in a blog post in January 2020 that these types of decision should not be made by private companies, *“which is why we advocate regulation that applies to the entire industry. In the absence of regulation, Facebook and other companies are free to choose their own policies”*.

It is certainly worth discussing whether Facebook’s interpretation of freedom of expression has merit from an ethical perspective. However, Rob Leathern’s statement draws attention to a specific question – namely the lack of, or at least incompleteness of, regulation.

[18] https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

[19] https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

[20] YouTube, on the other hand, another platform that contributes to the virality of false information through its recommendation algorithms, deleted the video but refused to make a clear statement about how it would handle deepfakes in the future.

[21] https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413

4.3 Regulation attempts by legislators

In Germany, deepfakes fall under *"general and abstract rules"* according to the response by the federal government to the brief parliamentary enquiry submitted by the FDP parliamentary group, as mentioned above. *"There are no specific regulations at the federal level that exclusively cover deepfake applications or were created for such applications. The federal government is constantly reviewing the legal framework at the federal level to determine whether any adjustment is necessary to address technological or social challenges."*

This means that some partial aspects of the deepfake issue, including revenge pornography, are supposedly implicitly covered by existing laws, but there is in fact no explicit approach to handling manipulated content. This applies to the entire spectrum of disinformation in digital space, not just the special case of *"deepfakes"*. As noted by the author of the study *"Regulatory responses to disinformation"* [22] from Stiftung Neue Verantwortung: *"previous attempts at regulation and political solutions [in Germany and Europe] are hardly suitable to curb disinformation."* A study by the law firm WilmerHale, *"Deepfake Legislation: A Nationwide Survey"*, [23] gives a detailed analysis of the status of deepfake regulation in the US.

In the United States, explicit pieces of legislation on deepfakes have already been written into criminal law – for example in Virginia, where non-consensual deepfake pornography is punishable, and in Texas, where any deepfakes intended to influence voters are punishable. Similar legislation was also passed in California in September 2019.

Possibly the most in-depth regulation of deepfakes was undertaken by the Chinese legislators in late 2019. Chinese law requires the providers and users of online video messaging and audio information services to clearly mark all content that was created or modified using new technologies such as artificial intelligence. Although it is certainly worth considering whether similar regulations could also be adopted by other countries, the case of China leaves a bad aftertaste: the Chinese government itself uses technology-based disinformation to target protesters in Hong Kong, among other things, and it seems inevitable that these new regulations will be used as a pretext for further censorship.

Effectively regulating new technological phenomena is certainly not easy. It has often proved difficult in the past. To drive a car in 19th century England, for example, a second person was required to walk in front of the vehicle waving a red flag under the Locomotive Act of 1865. [24] Nevertheless, there are measures that legislators can already take to counteract the phenomenon of deepfakes. Since 96% of deepfakes are currently non-consensual pornography, it would be a good start to explicitly make this punishable, as has been done in Virginia and California. Regulating defamation, fraud and privacy rights can be handled similarly. Furthermore, legislators should create clear guidelines for digital platforms to handle deepfakes in particular and disinformation in general in a uniform manner.

These measures can range from labelling deepfakes as such and limiting their distribution (excluding them from recommendation algorithms) to deleting them. Promoting media literacy should also be made a priority for all citizens, regardless of age. An adequate understanding of how deepfakes are created and disseminated should enable citizens to recognise disinformation and avoid being misled.

[22] https://www.stiftung-nv.de/sites/default/files/regulatorische_reaktionen_auf_desinformation.pdf

[23] Matthew Ferraro, WilmerHale | Deepfake Legislation: A Nationwide Survey – State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media.

[24] https://sites.google.com/site/motormiscellany/motoring/law-and-the-motorist/locomotive-act-1865/

4.4 The responsibility of the individual: critical thinking and media literacy

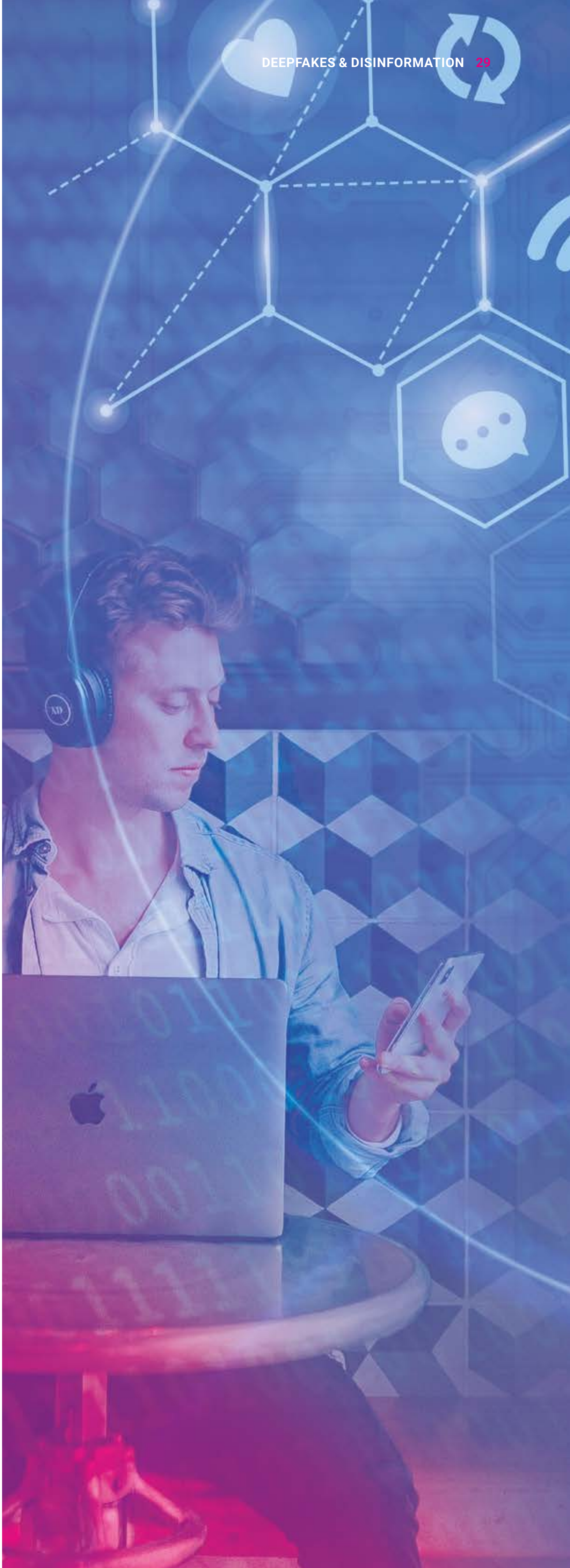
Critical thinking and media literacy are the basis for a differentiated approach to disinformation. It is certainly not possible and likely not desirable to ask every single person to question everything they see.

But more than ever before, people would be well advised to consume online content with caution. The simplest thing that anyone can do if an image, video or text seems suspicious is a Google search. Often, this will quickly unmask manipulated content, since the details of the manipulation circulate just as quickly as the content itself.

This is especially important for users who wish to share the content by *"liking it"* or commenting on it. We can also pay more attention to whether the blinking, facial expressions or speech in a video appear unnatural, whether parts of an image are blurred, or whether objects seem out of place.

However, these clues will quickly disappear as deepfake technology advances. In the future, there could conceivably be browser add-ons that automatically identify manipulated content and notify users, similar to an ad blocker. But this requires us to be aware of the possibility of manipulated content in the first place.

To raise this kind of awareness among its citizens, Finland, the country that was ranked the highest in a study measuring resilience to disinformation,[25] offers educational opportunities to its entire population – from kindergarten to retirement age.



[25] https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf

5.0 WHAT’S NEXT?

It is not yet possible to accurately predict the extent of the concrete effect that deepfakes will have on politics and society, but this does not excuse inaction. As highlighted above, neither fake videos nor disinformation are a new phenomenon as such – the novelty is the increasing simplicity of creating such content, its constantly improving quality and its capacity to be disseminated.

The presidential elections in the United States in autumn 2020 will undoubtedly prove a good litmus test. Nevertheless, the recommendation here cannot just be *“wait and see”*.

Researchers, technology companies, journalists, governments, and users themselves should make every effort to neutralise the negative impact of fake content. The first step is to implement explicit regulation and strong countermeasures against deepfake pornography, since this is already a widespread phenomenon that causes significant harm to its, mainly female, victims.

Uniform legal regulations on handling manipulated content in the media and on social media platforms are also required. We should not leave it to Facebook, Twitter, YouTube and other companies to decide what content falls under freedom of expression and what goes beyond it.

This task is the responsibility of legislators and constitutional democracy. However, we should not give in to the temptation to ban deepfakes completely. Besides its risks, the technology opens up promising new opportunities – in education, film and satire, among other things. Technology itself is neutral – it is people who use the technology to either benefit or harm society.

Author

Agnieszka M. Walorska

Agnieszka M. Walorska is a digitisation expert and Executive Director at the management and technology consultancy Capco. She has led several transformation and innovation projects for banks, insurance companies, and pharmaceuticall and automotive companies, among others.

She founded the digital strategy consultancy CREATIVE CONSTRUCTION, which was acquired by Capco in 2020. She is particularly interested in artificial intelligence and its impact on human-machine interaction and thus on business models and society.

With the Digital Innovation Breakfast, she has created a series of events with prominent speakers on these topics. She has published numerous studies and articles and regularly speaks at conferences and in companies.

She has contributed a book chapter on the algorithmic society, in which she deals with the ethical issues of artificial intelligence. The corresponding volume was published by the scientific publisher Springer in March 2020.

She studied social and political sciences at the University of Warsaw and Humboldt University Berlin and was a scholarship holder of the Hertie Foundation and the German Academic Scholarship Foundation.

freiheit.org