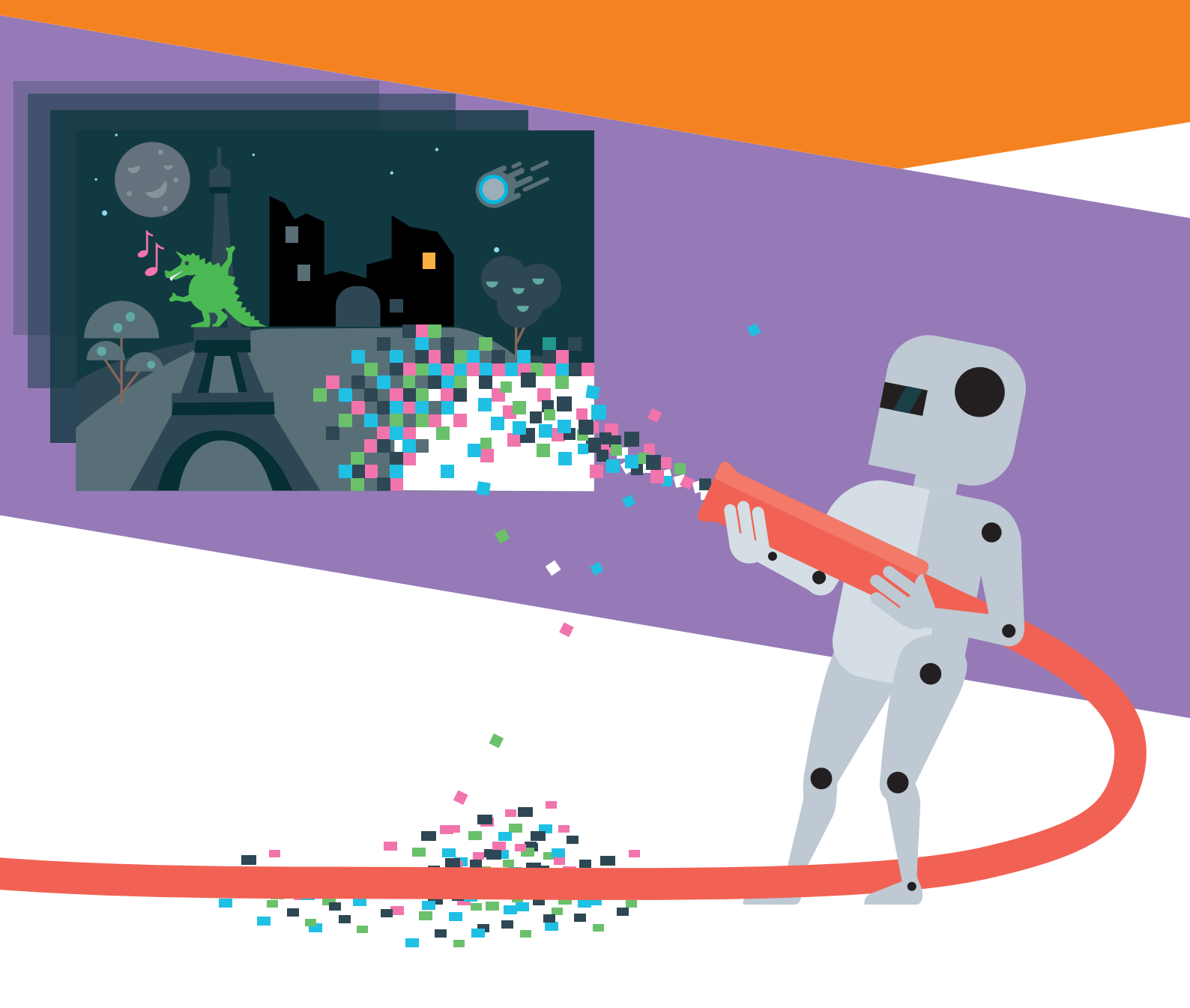
Eyes and Ears on the Test Bench

Condensed version of the TA-SWISS study, "Deepfakes and Manipulated Realities"



TA-SWISS, the Foundation for Technology Assessment and a centre for excellence of the Swiss Academies of Arts and Sciences, deals with the opportunities and risks of new technologies.

This abridged version is based on a scientific study carried out on behalf of TA-SWISS by an interdisciplinary project team under the leadership of Murat Karaboga (Fraunhofer Institute for Systems and Innovation Research, Karlsruhe, Germany). The abridged version presents the most important results and conclusions of the study in condensed form and is aimed at a broad audience.

Team members: Nula Frei (Institute for European Law, University of Fribourg in Uechtland), Manuel Puppis and Patric Raemy (Dept. For Communication and Media Research, University of Fribourg in Uechtland), Daniel Vogler (Research Centre for the Public Sphere and Society, University of Zurich), Frank Ebbers (Competence Centre for Emerging Technologies, Institute for Systems and Innovation Research, Karlsruhe, Germany), Greta Runge (Fraunhofer Institute for Systems and Innovation Research, Karlsruhe, Germany), Adrian Rauchfleisch (Graduate Institute of Journalism, Taiwan University), Gabriele de Seta (Department for Linguistics, Literature and Aesthetic Studies, University of Bergen, Norway), Gwendolyn Gurr (Audience Data Analyst, Swiss Radio and Television), Michael Friedewald (Fraunhofer Institute for Systems and Innovation Research, Karlsruhe, Germany), Sophia Rovelli (Institute for European Law, University of Fribourg in Uechtland).

Condensed version of the TA-SWISS study, "Deepfakes and Manipulated Realities"

Murat Karaboga, Nula Frei, Manuel Puppis, Daniel Vogler, Patric Raemy, Frank Ebbers, Greta Runge, Adrian Rauchfleisch, Gabriele de Seta, Gwendolyn Gurr, Michael Friedewald, Sophia Rovelli

TA-SWISS, Foundation for Technology Assessment (Ed.)
vdf Hochschulverlag an der ETH Zürich, 2024.

ISBN: 978-3-7281-4185-9

Also available in open access: www.vdf.ch

This abridged version can be downloaded at no cost: www.ta-swiss.ch

Deepfakes in a nutshell 4
Some opportunities … 4
… and risks 4
Urgent recommendations 5

Distorted view of reality 5
Profound deception 5
Pioneering work in the field of pornography 6
Competing for the best fake 6
Puppetry 6
Synthetic voices 7
Instruments for identifying deepfakes 8
Detecting signs of falsification 8
Need for healthy scepticism 9

How the general public and media representatives perceive deepfakes 10
Greater threat to society than to individuals 10
Perception of opportunities influenced by the applied designation 10
Tips of little assistance, but familiarity with new media helpful 10
Challenges for journalism 11
Moderate concern in Swiss editorial offices 12
Reliable sources as the antithesis of deepfake dissemination 12
Differing legal requirements for journalistic media and online platforms 12

When avatars interfere in politics and the economy 13
Humour as an election campaign aid 13
Call for increased vigilance at the political level 13
Potentials for entertainment and education 13
Industrial espionage with stolen identities 14
Switzerland as an attractive target 15

Deepfakes in the eyes of the law 15
Copyright protection for creative achievements 15
Limits of freedom of information 15
Identity theft, reputational damage and fraud 16
Sophisticated falsification of documents 16
Synthetic media as a prosecution instrument 16
International cooperation in the fight against globalised crimes 16

Correcting distortions of reality: some recommendations for dealing with deepfakes 18
Personal responsibility 18
Requirements on platforms, better protection for victims 18
Use of advanced technologies to defend against deepfakes 19
Information about the risks and opportunities 19

Deepfakes in a nutshell

The development of artificial intelligence (AI) is proceeding at a rapid pace, as is the production of "synthetic" videos, images and audio recordings. These do not present any genuine facts – they are created by computer software. Because it is becoming easier to produce deepfakes, it is to be expected that their social significance will increase rapidly. Deepfakes open up opportunities in the entertainment sector, and in the field of training and further education. But the associated risks – above all in the political sphere, in the context of mobbing of individuals and from the point of view of financial crime – cannot be dismissed out of hand.

Deepfakes – or synthetic media – are photos, videos or audio recordings that are created by means of artificial intelligence and present content that has never existed in the submitted form. They may take the form of manipulated or entirely synthetic files generated by software that is based on training data from enormous inventories on the Internet. The currently established deepfake programs encompass a broad variety of tools, ranging from easy-to-use software for substituting people's faces (face swapping), through to sophisticated applications for "virtual puppetry" with synthetic humans. In addition, various programs are already available that can generate videos – though for the time being only in rudimentary form – on the basis of prompts.

Some opportunities …

Synthetic media offer positive potentials for the entertainment industry. They could also open up certain opportunities for other business sectors – for example, when virtual influencers promote clothing or other products. In schools, history lessons could be rendered more attractive if avatars of figures from the past – for example, Caesar, Catherine the Great, Napoleon – were to chat interactively with students. Investigatory authorities also anticipate potential benefits from the possibility of visualising the sequence of events when investigating a crime.

… and risks

Deepfakes can be misused in order to depict people committing inadmissible acts that they did not carry out, or to put words in their mouth that they did not say. Videos and audio recordings of this type can be used to blackmail or compromise people – something that can occur in political disputes. They can also wreak havoc in private relationships, for example in the form of faked revenge porn.

People's voices can be cloned in order to fraudulently obtain money from their family or friends. And cloned voices of company managers can be misused for other financial crimes, for example to obtain business secrets.

For the media, deepfakes pose significant challenges because they face the time-consuming task of verifying videos so that they do not inadvertently distribute deepfakes themselves.

Urgent recommendations

In view of the rapid technological developments, a combination of safeguards is required in order to ensure that the positive potentials of synthetic media can be fully utilised and harmful impacts can be kept within bounds. Political measures, deepfake detectors, the designation of synthetic media as envisioned by major software providers, and sensitisation by the media to deepfakes, have to complement one another. In addition, it is important for everyone to assume personal responsibility. Internet videos should be viewed with a healthy degree of scepticism, and private images and videos should be uploaded with caution.

The relevant authorities should force large online platforms to delete deepfakes that harm individuals. And because in most cases individuals lose out in conflicts with large online platforms, there is a need for specialist agencies that can advise and support victims of deepfakes or those subjected to unjustified deletions. Advisory centres for victims of cyber crimes should be provided with sufficient funding by the federal government and the cantons.

Distorted view of reality

Generally speaking, what we see with our eyes and hear with our ears we believe to be true. In particular, we seldom doubt that video clips depict reality. Or at least this used to be the case until a few years ago. In the meantime, however, with new technologies and little effort it is possible to produce deceptively realistic videos and audio recordings of events that never took place.

In autumn 2023, for a brief time it appeared that a future super model was about to be discovered: visitors to Emily Pellegrini's new Instagram channel went into raptures over the videos posted there, and the number of her followers increased rapidly. And in addition to the huge number of heart on fire emojis, countless contact requests appeared in the comments column. Newspapers reported that a German footballer repeatedly asked for a date with her, and a billionaire, a tennis star and other leading figures from the world of sport also showed an interest in her until they had to acknowledge the fact that it was not a woman of flesh and blood that had attracted them, but rather an avatar generated with the aid of artificial intelligence. Her creator – who remained anonymous – announced that the woman of countless men's dreams served as a role model for him, and according to the British newspaper, Daily Mail, the artificial Emily brought him revenue of 10,000 dollars a month.

Emily Pellegrini, the (in her own words) "fun-loving girl", whose physical assets are presented on subscription service platforms such as "Onlyfans" and "Fanvue", represents a new type of influencer: synthetically created, but deceptively realistic (usually female) figures, who can even chat with their viewers thanks to AI-based text generators. The advantages that fake persons can bring to advertising are obvious: once they have been created they do not demand an hourly wage, never grow tired and always faithfully carry out instructions.

Profound deception

Using software based on artificial intelligence or artificial neural networks it is possible to produce videos that present content that in reality never existed. This may take the form of footage of natural disasters or explosions that have never occurred, or videos of celebrities saying or doing something they have never said or done. For example, a short film created by the Amsterdam video artist Bob de Jong depicts the former Prime Minister of the Netherlands, Mark Rutte, with a quavering double chin while soulfully playing the solo part from "Silent Night" on a violin. Bob de Jong uploaded this video to his YouTube channel, "Diep Nep", or "Deepfake". This term has also become established in other languages to designate videos and images that appear to be authentic, but have in fact been heavily manipulated or entirely generated on a computer. Experts also use the term, "synthetic media". The required material comprises data taken from the Internet, particularly images, videos and audio recordings from social media and video platforms. Whereas

Bob de Jong openly declares his creations as artefacts, in many – if not almost all – cases, the authorship of deepfakes remains unknown. The terms "deepfake" and "synthetic media" are used synonymously in this condensed version of the report.

Pioneering work in the field of pornography

Fake videos first appeared in autumn 2017 on Reddit, a kind of electronic repository for content from social media sites. These video clips were uploaded by a user under the name "DeepFake", who had exchanged the faces of the original actresses in pornographic videos with those of Emma Watson, Gal Gadot and other film stars. Shortly thereafter, another Reddit user provided a software called "FakeApp" which enabled everyone to produce their own deepfakes. Suddenly, what had previously been the domain of well-resourced Hollywood studios – namely the highly complex production of 3D computer graphics – was now available to everyone who was able to implement the five basic production steps specified by FakeApp.

Subsequently, vast quantities of manipulated porn videos were uploaded to the dark web. Initially, due to their low resolution and jerky images, these videos were easily identifiable as fakes, but with rapidly improving technologies they soon appeared to be increasingly authentic. In the meantime, it is not only images of prominent figures that are being misused for such purposes. Today, anyone can become a victim who, for example, may have annoyed someone. In February 2018, Wikipedia allocated a dedicated page to the term "revenge porn", and also cited the production of faked nude videos.

It is estimated that a large proportion of uploaded deepfakes depict pornographic images of women – although the genre has meanwhile moved into other areas, most notably politics. Large quantities of images of well-known personalities that circulate on the Internet provide a great deal of material that can be used for creating fake videos.

Competing for the best fake

A technology called generative adversarial networks (GANs) is attracting a great deal of attention in the field of production of deepfakes. GAN is the name given to computer programs that can generate new data with characteristics similar to those of a training set. The programs comprise two components – a generator and a discriminator – that compete against each other. While the generator attempts to produce similar images to those from the training set, the discriminator sets out to determine differences between the newly generated images and the training data. When competing against each other, the generator and discriminator mutually enhance one another so that, ultimately, images are created that stylistically can practically no longer be distinguished from the genuine original.

The alternative approach consists in the application of autoencoders. An autoencoder is an artificial neural network that is able to filter out essential characteristics from an image dataset and transfer them to other images.

Puppetry

The manipulations effected in deepfakes reach into the box of tricks to differing degrees – though this does not reveal anything about the respective effects on the public.

Manipulations can be limited to changing a person's facial expressions and mouth movements. Here, the facial expression of an actor can be transferred to the target person. Specialists refer to this process as facial re-enactment. It can also be applied if people synchronise themselves in a video, i.e. if their mouth movements are coordinated with their speech in another language. For example, a company called HeyGen Labs has developed an AI-based video program that records sound from a video, translates the content and re-transfers the new content into the video. Here the speaker's voice is cloned and the person's lips move in synchronisation with the voice content. In this way, journalists, presenters, etc. can synchronise their spoken content themselves.

Face morphing is another type of deepfake. Here the facial features of two individuals are merged into one. This technology is primarily used in criminal circles in order to falsify identification documents so that they can be used by several people at the same time.

Face swapping is a technology that has existed since the early phase of deepfakes described above. It was initially used for the production of faked porn videos. It is however also used playfully. Free apps are circulating on the Internet that enable users to modify iconic movie scenes, for example by substituting the face of Leonardo di Caprio or Kate Winslet with their own portrait.

With AI-based software it is also possible to construct new images of people from scratch who do not exist at all. Images and videos generated by AI face generators can be used as avatars in video games or as virtual dialogue partners in fully-automated customer support services.

The software acts as a puppeteer when it alters the poses and movements of a person in a video. This type of deepfake, which is called full body puppetry, is regarded as the most complicated. Currently, no integral AI package is available that would be capable of producing a complete video with language animation and voice synthesisation. Instead, several often expensive programs that are also complicated to use have to be combined, and this makes the creation of deceptively realistic deepfake videos a difficult undertaking.

The production of synthetic videos at the push of a button is likely to be within easy reach in the near future, however: based on the model of AI-supported image generators, which can create a photo-realistic picture of, for example, a unicorn or a forgery of a Rembrandt painting, with an initial series of AI-supported programs it is possible to generate deepfake videos using voice or text commands. This type of software will probably be available to a broader circle of users in the foreseeable future.

Synthetic voices

It is not only our eyes, but also our ears, that can be deceived – by software that is able to clone a person's voice and speaking habits. In 2018, an experiment carried out by a Scottish company hit the headlines when it succeeded in reanimating the voice of the assassinated US President, John F. Kennedy: based on the original manuscript and numerous voice recordings of Kennedy, the firm succeeded in generating an audio playback of the speech the President was unable to hold in autumn 1963 due to his assassination. Both his Boston accent and his cadence were imitated to perfection.

In the meantime, this technology has been further developed. In order to generate a model of a person's way of speaking, all that is required is a standard laptop and a few seconds of an audio clip – for example, recorded from a lecture uploaded to YouTube. Progress has also been made in the field of speech synthesis. This is based on software that converts the written word into an audio clip. This technology is used for the automatic production of audio books, as well as by visually impaired people when they want to have texts read aloud to them.

Instruments for identifying deepfakes

The somewhat troubling term, "deepfake", indicates the objective that is pursued by many synthetic videos and audio recordings: to mislead people and influence them to the benefit of the originator. Experts are discussing various ways of dealing with audio and visual deceptions.

One possible solution would be to require transparency with respect to the origin of a video or audio recording. This could take the form of a digital signature that would have to be integrated into the video or audio file during its production. Based on the block chain this would be relatively easy to implement, but with considerable time and effort it would also be possible to falsify this type of digital fingerprint. In addition, experts warn that authoritarian regimes and secret services would have an instrument at their disposal for tracking down whistle-blowers, human rights activists, undesirable journalists, etc. Furthermore, a signature could certify the origin of a recording, but not whether the whole content or only a portion thereof was recorded. It would also not be possible to identify videos that depict events re-enacted with actors. While such videos would be technically "genuine", they could nonetheless distribute falsehoods. A Norwegian TV series captures such a scenario in which a partisan group shows a video of the – alleged – assassination of a controversial prime minister so that he can more easily go into hiding.

Detecting signs of falsification

Other methods are intended to unmask artificial videos and audio recordings on the basis of certain characteristics. However, broad agreement exists regarding cloned voices: these acoustic deepfakes now sound so genuine that it is barely possible to distinguish them from the real voice of the person concerned, especially when the recording is played over the telephone. Police have issued warnings about the latest grandparent scams, which are now likely to be even more successful because the targeted victims believe they are hearing the voice of a relative asking for money.

By contrast, fake videos can be detected through certain artefacts that cannot readily be recognised by the human eye, but can be spotted with the aid of AI-supported algorithms. Faulty edges, unnatural cross-fading, distortions and blurring can be signs of a deepfake. In the meantime, detector programs are available that are designed to expose faked videos. In the course of this study, TA-SWISS tested two free detectors. The findings were far from satisfactory, because both programs delivered false results. Both detectors also declared some genuine videos to be fake, which could undermine the trustworthiness of original content. In any case, it is to be anticipated that developers of deepfake software will also become aware of the tell-tale signs and do everything in their power to correspondingly improve their programs – a cat-and-mouse game in

which the creators of fake videos are likely to maintain the upper hand in the foreseeable future.

Need for healthy scepticism

Currently, common sense is probably on a par with detectors when it comes to identifying deepfakes. Questioning the source and content of a video, and paying close attention to details such as strands of hair, fingers, earrings, etc., and to any unusual behaviour by the filmed person, can help identify deepfakes. In addition, the ability to identify fake videos can be trained on websites such as “Detect-fakes”.

In any case, with a certain degree of scepticism the German footballer and other admirers of Emily Pellegrini might not have been taken in by that deepfake. When comparing videos, it becomes apparent that the proportions of the artificial influencer vary. And even the influencer’s cheerful to occasionally somewhat submissive responses in chats should be viewed with mistrust. A certain degree of scepticism is also called for in the case of surreal perfection or exaggerated glamour.

Methodology applied in this study

In the study on deepfakes conducted by TA-SWISS, in addition to comprehensive resarch of existing literature, several surveys were carried out. In an online survey supplemented by an online experiment, members of the general public were asked about their experiences and dealings with deepfakes. In addition, the project team interviewed and questioned media representatives, public administration personnel and politicians in order to learn how they assess the risks, opportunities and impacts associated with fake videos and audio recordings. The project team also tested the capability of several free deepfake detectors to identify fake videos.

How the general public and media representatives perceive deepfakes

In Switzerland, to date the general public have barely come into contact with deepfakes. The most likely places to encounter them are platforms such as YouTube, TikTok and Instagram. The majority of respondents associate deepfakes with risks, and are generally not able to distinguish well-made deepfake videos from genuine ones. In Switzerland's major media organisations too, deepfakes are primarily regarded as a risk. By addressing the topic of deepfakes, journalistic media play a significant role in sensitising the general public to the issue.

The study conducted by TA-SWISS is the first comprehensive analysis of the perception of deepfakes in Switzerland. Of the over 1,300 respondents, slightly more than half stated that they were familiar with the term "deepfake", and slightly fewer than half replied that they had already seen such a video. Only a small minority (around two to three percent) reported that they themselves had created or distributed deepfakes here. Overall, the findings of the TA-SWISS study show that people in Switzerland have very little experience with deepfake technologies. Here, typical influencing factors such as age, gender and education, which normally play a role when people encounter new technologies, are of little significance.

Greater threat to society than to individuals

Swiss citizens perceive deepfake technologies as more of a risk than an opportunity. Above all, people fear that fake news distributed in the form of deepfakes could undermine trust in the country's information media. The risk that deepfakes could influence popular votes and elections in Switzerland is considered to be somewhat less acute.

The respondents assess the risk of becoming personal victims of deepfakes as relatively low. It is evident that women regard this risk to be greater than men do, which in view of the numerous pornographic deepfakes hardly comes as a surprise.

Perception of opportunities influenced by the applied designation

When asked about the opportunities that could go hand in hand with deepfakes, the respondents express scepticism. But the situation is different when the more neutral term "synthetic media" is used instead of "deepfake". In a preliminary survey, the participants were divided into two groups that were given different questionnaires. One group received a questionnaire in which the term "deepfake" was used, while for the other group the term "synthetic media" was used for all questions.

The results show that the term "synthetic media" is less known: around two-thirds of the respondents are familiar with the term "deepfake", whereas only slightly over one-third understand what is meant by the term "synthetic media". The assessment of risk is more or less the same in both groups, but with respect to the perception of opportunities the situation is quite different: here the respondents regard the opportunities associated with synthetic media in terms of their impact on media and the economy to be significantly higher than those associated with deepfakes. Thus the way in which people perceive the benefits of a given technology ultimately depends on the applied designation.

Tips of little assistance, but familiarity with new media helpful

The findings of the TA-SWISS study indicate how difficult it is to recognise deepfakes. In an experiment, the respondents were shown three deepfake videos and three genuine ones, and were asked to assess the reality of their content. Here, too, the participants were divided into two groups, one of which received tips intended to help them identify deepfakes.

The outcome was that the participants felt unable to assess the content with any certainty, i.e. they were barely able to distinguish well-made deepfake videos from genuine ones. Furthermore, the respondents in the group that had received tips in advance for recognising deepfakes were not able to assess the videos any better than the group that was not given prior assistance.

On the other hand, it was clear that experience in the use of social media correlates positively with the recognition of deepfakes. Thus a high level of media literacy facilitates the identification of fake videos: it is important to not only be familiar with conventional mass media, but also to learn to treat information from unknown sources on social media with appropriate caution.

Challenges for journalism

Recognising fake information and disinformation is one of the core tasks of media representatives, for whom deceptively realistic videos and audio recordings pose additional challenges. Because journalism critically assesses political events and influences the formation of public opinion and decision-making, the correct recognition of deepfakes by media representatives is of relevance to society as a whole. But it is also in the own interests of the media to ensure that they themselves do not (unintentionally) disseminate deepfakes, because otherwise their credibility – and thus the business models of media organisations – could suffer severe reputational damage.

Media representatives have to quickly verify the authenticity of videos and audio recordings, but this is in fact a time-consuming process. In addition, many media organisations are under financial pressure, and not all of them are able to afford to hire the necessary specialised personnel. Furthermore, while press articles can sensitise the public to the problem of deepfakes, if the media draw a great deal of (or too much) attention to deepfakes in their reports, this could increase the risk that scepticism would become more widespread among the public, and could thus give rise to mistrust of media content in general.

The nature of their work means that journalists are often exposed to significant risks: as occurrences in India and the USA have shown, top journalists and reporters can themselves become victims of deepfakes. And even though they are generally not as prominent as some of their international colleagues, many Swiss journalists are also exposed to such risks. Deepfakes could thus be used as an additional intimidation instrument.

Moderate concern in Swiss editorial offices

The survey among Swiss media organisations carried out within the framework of the TA-SWISS study shows that while Swiss editorial offices are well aware of the phenomenon of deepfakes and the problem is also addressed in the education of journalists, the risk is not regarded as particularly high. Here, fake videos are treated as a subcategory of disinformation. Media representatives in Switzerland are not (yet) unduly concerned about becoming victims of deepfakes themselves.

Swiss editorial offices primarily find themselves confronted with fake videos in foreign reporting, for example in the context of the war in Ukraine. Here, the challenge for editors is to recognise fake videos so that they do not unknowingly disseminate them. In the survey, the opinion was expressed that Swiss editors could benefit from the fact that major foreign media organisations possess highly qualified research teams that can verify the authenticity of videos. Furthermore, Switzerland is not in the firing line of deepfake production facilities because its media attract little interest outside the country.

However, the survey among Swiss media organisations also revealed that especially complex cases place high demands on verification – and that careful verification in editorial offices does not suffice on its own, but needs to be supplemented by educating and sensitising the public. Verification by the media of the authenticity of the content of news reports is not sufficient. It is also necessary to ensure that everyone is aware of the problem of manipulated information.

Reliable sources as the antithesis of deepfake dissemination

The majority of the surveyed media representatives regard deepfakes as a risk. In their view, a certain potential exists at most in the personalisation of news services through synthetic presentation or the use of avatars in research.

The sole benefit that can be attributed to faked videos and images is that they could strengthen the position of journalistic media as reliable information sources, as long as the journalists succeed in clearly identifying deepfakes and other manipulations at an early stage and differentiate themselves from other less trustworthy sources.

Differing legal requirements for journalistic media and online platforms

With respect to the dissemination of fake videos, the legal requirements governing journalistic media differ from those that apply to online platforms. Provisions governing conventional media are laid down in the Swiss Federal Constitution, where Article 93 stipulates that radio and television broadcasts must present events objectively and adequately reflect diversity of opinion. The Federal Radio and Television Act also specifies minimum requirements on programme content, and stipulates that facts and events must be presented objectively so that the audience is able to form its own opinion. All presented opinions and comments must be clearly recognisable as such. There are also authorities to whom people can submit complaints if media do not comply with the requirement of objectivity and impartiality.

However, online platforms such as social networks and video-sharing services, which distribute content uploaded by their users, are not required to declare whether a posted video is a deepfake. On the contrary, content distributed on social media is protected by the right to freedom of expression, which means the authorities can only take action against evidently unlawful content. The uploading of pornographic videos is subject to the provisions of the Federal Act on the Protection of Minors in the Areas of Film and Video Games. This legislation obliges the distributors of such content – including streaming services – to implement measures to protect minors and where necessary to restrict access to videos of this nature. Enforcing Swiss legislation against foreign providers and services, however, is one of the biggest challenges in dealing with deepfakes.

When avatars interfere in politics and the economy

Whether in armed conflicts or during election campaigns, when the going gets tough, deepfakes are used in order to unsettle the opponent. And in the economy, faked films form an integral part of cyber crime.

It is hardly surprising that deepfakes are used during election campaigns in order to discredit political opponents and confuse the targeted public.

In March 2022, for example, a disseminated deepfake showed Ukrainian President Zelensky holding a speech in which he apparently urged the population to surrender to the invading Russian armed forces. And in Pakistan, at the beginning of 2024 the jailed leader of the opposition, Imran Khan, addressed his compatriots from his prison cell and participated in the election campaign with an AI clone of himself. In the USA, too, attempts were made to influence the election campaign through the use of a series of deepfakes. Here, for example, members of the Democratic Party in New Hampshire received a fake phone call from Joe Biden in January 2024: in this robo-call he asked the recipients to boycott the primary elections.

Synthetic media have also already been used in Switzerland in the political sphere. In summer 2023, a campaign poster entirely produced using AI, which prominently depicted an ambulance obstructed by climate activists, caused an uproar – but an incident of this nature never actually occurred here. And in October 2023, a faked video depicted National Councillor Sibel Arslan making an appeal that was in contradiction with her declared values and causes.

Humour as an election campaign aid

It is a fact that deepfakes can confuse the public and bring politically active figures into disrepute, intimidate them or in some cases elicit confidential information from them. But irrespective of their damaging effect, synthetically produced videos can also have positive political potential. With humorous content they could stimulate political debate and help people form an opinion. And deepfakes can also be used in order to explain complex issues to voters.

In addition, politicians could address voters using clearly designated deepfakes with satirical and amusing content. The use of humour and wit is an ideal way to reach a broader audience and thus attract the attention of the public.

Call for increased vigilance at the political level

The authors of the TA-SWISS study asked members of the Swiss parliament and employees of the federal administration how they perceive and assess deepfakes.

Faked videos have clearly been observed in daily political life, for the majority of the respondents stated that deepfakes are already an issue in their work. Asked whether faked videos should be assessed as a risk or an opportunity, their response was unanimous: fake videos are regarded as a risk, and no one identified any positive aspects.

Here the main concern was the perceived threat to Swiss democracy and to the confidence in local institutions. The respondents also cited as relevant risks the possibility that they themselves could become victims or protagonists of a deepfake, or be taken in by one, or even that a fake video could negatively affect international relations – although they felt that such occurrences were fairly unlikely. The respondents unanimously stated that inadequate measures were being taken to protect against deepfakes.

Potentials for entertainment and education

The effects of deepfakes appear to be less negative in the economy than in the political arena and the public administration. In the view of the entertainment sector they offer a variety of benefits, for example in the film industry. The gaming industry is looking for ways in which players' faces can be transferred to their avatars. And the advertising industry envisages benefits from the use of synthetic influencers who can present fashion wear, for example, or are active in the field of corporate communication.

Deepfakes can also assist with fundraising in campaigns conducted by charities. In 2019, a synthetic double of former football star David Beckham appealed in nine languages for people to sign a petition to combat malaria, and thus for the heads of state in the most seriously affected countries to do more to prevent the disease.

In classrooms, it would be possible to increase children's interest in history lessons if, for example, Cleopatra, Napoleon or other historical figures were to interactively chat with them in the form of deepfakes. In addition, avatars in personalised distance education courses could increase the motivation of children and youths to learn. There are also potential uses for synthetic media in the medical sector, for example in the field of treatment of anxiety disorders: here, an avatar of the patient can be placed in the situation that normally causes anxiety, for example balancing at a great height. From a psychological perspective this facilitates an objective confrontation with the situation concerned.

Industrial espionage with stolen identities

But deepfakes nonetheless represent certain risks for the economy – they can give rise to reputational damage in the same way as in the political sphere. Faked private statements by a presumed insider can ruin a company's reputation or manipulate the equity markets. In their turn, faked influencers can be used for advertising fraud and ultimately undermine trust in the provider.

In addition, deepfakes facilitate identity theft. With a cloned voice or a 3D avatar of a given person it is possible to outsmart voice or face recognition systems. In this way, criminals can access a private individual's account or unlawfully obtain business secrets.

Admittedly, deepfake attacks against players in the economy do not raise fundamentally new questions, and they are now an integral part of cyber crime. The perpetrators of cyber and deepfake crimes also share similar objectives: financial gain or sabotaging competitors. Companies and individuals who meet the criteria of high worth, high level of visibility, sluggish activity and relative ease of access are especially vulnerable.

Switzerland as an attractive target

As one of the world's most innovative and productive economies, Switzerland is regarded as an attractive target for cyber crime and thus deepfake attacks – especially because its security structure is lagging behind its status as a significant economy. On the 2020 global cybersecurity index, Switzerland was ranked 42nd of 182 countries. In a study carried out in the same year on behalf of the Federal Intelligence Service, 15 percent of the participating companies stated that they had already been a target of industrial espionage. However, very few of the targeted companies – just 13 percent – reported the incident to the police or the public prosecution service. In most cases, the companies concerned prefer to take internal measures – often backed up with external support. The fact that even flagship companies in the Swiss economy are not immune to cyber attacks was illustrated in 2023 when both the Neue Zürcher Zeitung (a major Swiss daily newspaper) and Swiss Federal Railways were victims of such attacks.

Since many companies prefer to keep a low profile and thus do not report cyber attacks, it is not possible to quantify the economic losses attributable to deepfakes, especially because the latter only account for a certain proportion of online attacks. By contrast, figures are available concerning the evolved illegal fake video market: in 2023, the costs for a basic fake video on the dark web were around 20 US dollars per minute. In addition to various services, users of the corresponding platforms also have access to tutorials, discussion posts and other aids relating to deepfakes. Tips for the distribution and sale of products and services relating to manipulated videos are primarily available in English and Russian in dark web forums, but darknet sites in Turkish, Spanish and Chinese are also highly active in this field.

Measures that could potentially be taken to reduce the risks associated with deepfakes are cited in the concluding chapter containing recommendations resulting from the TA-SWISS study.

Deepfakes in the eyes of the law

In Switzerland, the existing legislation covers most of the crimes that can be committed with deepfakes. However, many hurdles exist when it comes to enforcing the legal provisions, and there is an urgent need for international cooperation.

Artistic freedom and freedom of expression and information are fundamental rights on which Switzerland places a high value. These are guaranteed by the Swiss Federal Constitution as well as the European Convention for the Protection of Human Rights and Fundamental Freedoms. Deepfakes are also protected by fundamental rights, but their protection may be restricted if their fake content violates the rights of others.

Copyright protection for creative achievements

Synthetic media are protected by copyright as long as they can be classified as "works": this means that photographic, cinematographic and other visual and audiovisual products are protected by copyright law. The artistic or aesthetic content of a video or image is not of relevance here. It is the person's creative achievement that is decisive. But in the case of a fully-automated video produced by AI, this creative achievement is missing, and it is therefore questionable whether videos created without human intervention can be regarded as "works" and thus enjoy protection under copyright law or protection of artistic freedom.

Limits of freedom of information

Limits also apply to freedom of information. Deliberate false information does not fall within the scope of protection. But apart from the Federal Radio and Television Act, which obliges broadcasters to provide accurate information, there is no legislation that governs the veracity of the content of videos.

If threatening deepfakes terrorise the population, for example by frightening people with warnings of an impending disaster, the producers can be prosecuted. Anyone who terrorises the population by threatening or simulating a danger to life, health or property, can be prosecuted under Article 258 of the Swiss Penal Code.

Identity theft, reputational damage and fraud

The Swiss Civil Code protects a variety of personal rights, including the right to the use of one's own image, voice and name. Deepfakes that use photos and audio recordings of people without their consent constitute an infringement of their personal rights.

People who are depicted unfavourably in a deepfake can defend themselves: the Swiss Penal Code contains various articles governing libel and slander. In addition, people who are maliciously deceived, for example in order to get them to transfer money or disclose confidential information, can also call on the applicable articles of the Swiss Penal Code.

Criminal law also penalises the distribution of pornographic content – an offence that is probably committed by a large proportion of deepfakes. It is a criminal offence to publicly exhibit pornographic images and audio recordings or supply them to people without their consent. If such a deepfake is imposed on a private individual, this may constitute the offence of sexual harassment. A new article has been incorporated in the Swiss Penal Code that governs the phenomenon of revenge porn, which prohibits the unauthorised distribution of non-public sexual content.

Sophisticated falsification of documents

Recordings from surveillance cameras or bodycams are often used as evidence in court proceedings. Video recordings of this nature could be falsified with the aid of deepfake technologies, or vice versa, deepfakes could be used to create false alibis. The manipulation and falsification of documents is nothing new, but with synthetically generated videos and audio recordings an additional method is now available.

Due to the use of deepfakes, the evaluation of audiovisual evidence has become even more challenging. However, regardless of how procedural documents, videos and sound recordings are manipulated, anyone who falsifies a document is committing a crime. Similarly, anyone who deliberately falsely accuses someone can also be held accountable on the basis of false accusation or misdirection of the administration of justice.

Synthetic media as a prosecution instrument

Legal experts are discussing the possibility of using synthetic media as a prosecution instrument. In covert investigations, it is often necessary to upload child pornography material in order to infiltrate targeted online sites. However, the investigators themselves are prohibited from committing any offences when hunting criminals – and in Switzerland, the distribution of "fictitious" child pornography is a crime. Furthermore, the production of a fully synthetic child pornography deepfake is precarious, because producing an apparently genuine video requires images of actual child abuse. This means that the police cannot use any child pornography deepfakes for their investigations.

It would be conceivable to use deepfakes in criminal prosecution, for example in order to reconstruct the progress of a crime using mobile phone videos, data from surveillance cameras and bodycams. But this process, too, raises a variety of questions. Here, one problem concerns the fact that, while such a reconstruction may appear to be objective, it is based solely on the assumptions of the prosecution. It is also not clear how the right of the accused to participate in all procedural steps, including the "virtual crime scene investigation", can be guaranteed, and in what form the digitally collected evidence can be subsequently archived.

International cooperation in the fight against globalised crimes

Swiss legislation thus covers the majority of crimes that can be committed using deepfakes.

However, the enforcement of the legal provisions is likely to often encounter hurdles. For example, in most cases it is difficult to identify the authorship of deepfakes. And even if it proves possible to find the perpetrators, this is of little use if numerous copies of the deepfake have already been made and widely distributed. The majority of deepfakes are produced abroad and uploaded to platforms outside Switzerland. And prosecuting crimes committed in the Internet is usually extremely costly and time-consuming. It is often the case that several persons are involved who have to be identified, and on top of this there is the problem of unclear jurisdictions and overburdened prosecution authorities.

Through the implementation of judicial assistance agreements and intensified international cooperation in the area of data exchange, experts anticipate improvements with respect to cross-border enforcement. In the European Union, the aim of the Digital Services Act is to provide better protection for Internet users. One of its provisions stipulates that platforms must combat illegal content. They must also enable users to report such content and are required to cooperate with designated trusted flaggers. The term “trusted flaggers” refers to entities that detect unlawful content and notify the platform concerned. The EU also recently adopted legislation on artificial intelligence that includes a transparency requirement for deepfakes.

Online platforms and social networks have also been active: a number of them have formulated community guidelines that prohibit digital fakes and misleading information. In addition, 34 major companies – including Meta, Google, Microsoft and TikTok – have signed a code of conduct which obliges them to combat disinformation. But relying solely on self-regulation by the major platforms would hardly suffice to meet the public interest. With respect to the specification of deletion criteria there is a fundamental lack of democratic involvement and transparency, and the danger of unilateral exercising of power cannot be ruled out.

Correcting distortions of reality: some recommendations for dealing with deepfakes

Undesirable consequences of deepfakes cannot be prevented or even mitigated through regulatory or specific technical measures alone. Instead, what is required is a combination of precautionary measures at various levels and a high degree of personal responsibility so that it is also possible to benefit from the potentials of synthetic media.

Since the majority of manipulated videos reach their viewers via the major online platforms, the latter have a key role to play in the regulation of deepfakes. In addition, the relevant authorities, the communications industry, the education sector and ultimately the general public also have to play a role.

Personal responsibility

Training and further education in all areas of media and information literacy should figure at the top of the list of priorities in all sectors. And in their turn, all citizens should autonomously take advantage of the education opportunities and information services provided by various institutions. Personal responsibility is also essential when it comes to the evaluation and distribution, as well as the production, of deepfakes. In addition, everyone has to be made aware that uploading images and voice recordings can facilitate the production of deepfakes. The principle that the Internet never forgets applies especially with respect to deepfakes.

Anyone who likes to watch videos on the Internet or receives videos via social media should always be aware of the possibility that any video could be a fake. Scepticism is particularly called for if a video or voice recording is emotionally charged or particularly spectacular.

Requirements on platforms, better protection for victims

The relevant public authorities should take the necessary action to oblige platforms to delete or block reported deepfakes. Furthermore, platform operators should be required to set up a system for reporting deepfakes. Transparency requirements and objection options would strengthen the rights of victims of deepfakes, as well as of those affected by unjustified deletions. In order to enforce such measures, international cooperation is essential. This means that additional cooperation instruments have to be created at the international level. Switzerland should also campaign for the introduction of internationally applicable standards and regulations against cyber crime.

In disputes with major online platforms, private individuals usually have little chance of success. In view of this, there is a need for specialised agencies that advise and support victims of deepfakes, or those affected by unjustified deletions. Advisory centres for victims of cyber crimes should be provided with sufficient funding and personnel by the federal government and the cantons. Switzerland should officially recognise trusted flaggers, so that their reports of problematic deepfakes in the Internet would have to be given priority. The possibility of providing financial support for such informers should also be considered.

Use of advanced technologies to defend against deepfakes

Broad-based debate on authentication and identification processes would be a welcome step. Advanced methods, including multiple-factor authentication, could thwart attempts at deception using voice or face deepfakes. Here it is important to resort to the most advanced authentication methods wherever possible, because cyber criminals are also constantly working on ways to bypass security measures.

In view of the rapid development of deepfake technologies, every possible means should be applied to prevent misuse. Even tools such as deepfake detectors, which are not yet especially effective, can contribute to the range of tools for defending against deepfakes. As robust a structure as possible of existing safeguards is also recommended.

Information about the risks and opportunities

Currently, very few people have had much experience with deepfakes, and many know very little about them. Information provided in classrooms and the media should increase the degree of awareness of the phenomenon. And tips as to how to verify sources and question the plausibility of videos would also be useful. Schools should examine whether discussion about deepfakes could be incorporated into Switzerland's Syllabus 21 in order to foster media literacy.

Despite the risks that need to be clarified, the potential opportunities relating to synthetic videos should not be suppressed. Here it is important to observe the utilised wording, because people associate the term "deepfake" with opportunities to a much lesser extent than they do when a more neutral term such as "synthetic media" is used.

Advisory group

- **Prof. Dr. Reinhard Riedl**, Berner Fachhochschule BFH, president of the advisory group, member of the TA-SWISS Steering Committee
- **Dr. Bruno Baeriswyl**, data protection expert, president of the TA-SWISS Steering Committee
- **Cornelia Diethelm**, Centre for Digital Responsibility
- **Prof. Dr. Rainer Greifeneder**, Abteilung Sozialpsychologie, Universität Basel
- **Thomas Häussler**, Abteilung Medien / Sektion Grundlagen Medien, Federal Office of Communications OFCOM
- **Andrea Hauser**, Cybersecurity expert, Scip
- **Erich Herzog**, lawyer, member of the Executive Board Economiesuisse
- **Prof. Dr. Selina Ingold**, IDEE Institut für Innovation, Design & Engineering, Ostschweizer Fachhochschule
- **Melanie Kömle Bender**, media documentalist, Schweizer Radio und Fernsehen SRF
- **Thomas Müller**, science journalist, member of the TA-SWISS Steering Committee
- **Prof. Dr. René Schumann**, HES-SO Valais-Wallis, Forschungsinstitut Informatik
- **Prof. Dr. Giatgen Spinas**, Universität Zürich, member of the TA-SWISS Steering Committee
- **Dr. Stefan Vannoni**, economist, CEO cemsuisse, member of the TA-SWISS Steering Committee

Project management at TA-SWISS

- **Dr Elisabeth Ehrensperger**, Managing director
- **Dr Laetitia Ramelet**, Project manager
- **Dr Lucienne Rey**, Project manager

Impressum

Eyes and Ears on the Test Bench
Abridged version of the study “Deepfakes and Manipulated Realities”
TA-SWISS, Bern 2024
TA 81A/2024

Author: Lucienne Rey
Translation: Keith Hewlett
Production: Laetitia Ramelet and Fabian Schluep, TA-SWISS
Layout and graphics: Hannes Saxer, Bern
Printed by: Jordi AG – Das Medienhaus, Belp

TA-SWISS – Foundation for Technology Assessment

New technology often leads to decisive improvements in the quality of our lives. At the same time, however, it involves new types of risks whose consequences are not always predictable. The Foundation for Technology Assessment TA-SWISS examines the potential advantages and risks of new technological developments in the fields of life sciences and medicine, digitalisation and society as well as energy and environment. The studies carried out by the Foundation are aimed at the decision-making bodies in politics and the economy, as well as at the general public. In addition, TA-SWISS promotes the exchange of information and opinions between specialists in science, economics and politics and the public at large through participatory processes. Studies conducted and commissioned by the Foundation are aimed at providing objective, independent, and broad-based information on the advantages and risks of new technologies. To this purpose the studies are conducted in collaboration with groups comprised of experts in the relevant fields. The professional expertise of the supervisory groups covers a broad range of aspects of the issue under study.

The Fondation TA-SWISS is a centre for excellence of the Swiss Academies of Arts and Sciences.

TA-SWISS
Foundation for Technology Assessment
Brunngasse 36
CH-3011 Bern
info@ta-swiss.ch
www.ta-swiss.ch

member of the
swiss academies
of arts and sciences