ETSI GR SAI 011 V1.1.1 (2023-06)

Securing Artificial Intelligence (SAI); Automated Manipulation of Multimedia Identity Representations

Disclaimer

The present document has been produced and approved by the Securing Artificial Intelligence (SAI) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/SAI-0011

Keywords
artificial intelligence, identity

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Securing Artificial Intelligence (SAI).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document covers AI-based techniques for automatically manipulating existing or creating fake identity data represented in different media formats, such as audio, video and text (deepfakes). The present document describes the different technical approaches and analyses the threats posed by deepfakes in different attack scenarios. It then provides technical and organizational measures to mitigate these threats and discusses their effectiveness and limitations.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] Reuters, 2020: "Fact check: "Drunk" Nancy Pelosi video is manipulated".

[i.2] Karras et al., 2019: "Analyzing and Improving the Image Quality of StyleGAN".

[i.3] Gu et al., 2021: "StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis".

[i.4] Abdal et al., 2020: "StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows".

[i.5] Roich et al., 2021: "Pivotal Tuning for Latent-based Editing of Real Images".

[i.6] Zhang et al., 2020: "MIPGAN - Generating Robust and High Quality Morph Attacks Using Identity Prior Driven GAN".

[i.7] Tan et al., 2021: "A Survey on Neural Speech Synthesis".

[i.8] Qian et al., 2020: "Unsupervised Speech Decomposition via Triple Information Bottleneck".

[i.9] Casanova et al., 2021: "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone".

[i.10] VICE, 2017: "AI-Assisted porn has arrived - and Gal Gadot has been made its victim".

[i.11] NYTimes, 2020: "Deepfake Technology Enters the Documentary World".

[i.12] BuzzFeedVideo, 2018: "You Won't Believe What Obama Says In This Video!".

[i.13] C. Chan et al., 2019: "Everybody Dance Now".

[i.14] Adobe®, 2021: "Roto Brush and Refine Matte".

[i.15] Prajwal et al., 2020: "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild".

[i.16] Fried et al., 2019: "Text-based Editing of Talking-head Video".

[i.17] Zhou et al., 2021: "Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation".

[i.18] Hwang, 2020: "Deepfakes - A grounded threat assessment", Center for Security and Emerging Technology.

[i.19] Reuters, 2022: "Deepfake footage purports to show Ukrainian president capitulating".

[i.20] Forbes, 2021: "Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find".

[i.21] Forbes, 2019: "Deepfakes, Revenge Porn, And The Impact On Women".

[i.22] Shazeer Vaswani et al., 2017: "Attention is all you need". Advances in neural information processing systems, 30, pp.

[i.23] Irene Solaiman et al., 2019: "Release Strategies and the Social Impacts of Language Models".

[i.24] Vincenzo Ciancaglini et al., 2020: "Malicious Uses and Abuses of Artificial Intelligence", Trend Micro Research.

[i.25] Eugene Lim, Glencie Tan, Tan Kee Hock, 2021: "Hacking Humans with AI as a Service", DEF CON 29.

[i.26] Susan Zhang, 2022: "OPT: Open Pre-trained Transformer Language Models".

[i.27] Karen Hao, 2021: "The race to understand the exhilarating, dangerous world of language AI", MIT Technology Review.

[i.28] Ben Buchanan et al., 2021: "Truth, Lies, and Automation How Language Models Could Change Disinformation", Center for Security and Emerging Technology.

[i.29] Cooper Raterink, 2021: "Assessing the risks of language model "deepfakes" to democracy".

[i.30] Li Dong et al., 2019: "Unified Language Model Pre-training for Natural Language Understanding and Generation", Advances in Neural Information Processing Systems, Curran Associates, Inc.

[i.31] Almira Osmanovic Thunström: "We Asked GPT-3 to Write an Academic Paper about Itself-Then We Tried to Get It Published".

[i.32] Tom B. Brown et al, 2020: "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Curran Associates, Inc.

[i.33] OpenAI, 2019: "Better Language Models and Their Implications".

[i.34] David M. J. Lazer et al., 2018: "The science of fake news".

[i.35] Mark Chen et al., 2021: "Evaluating Large Language Models Trained on Code".

[i.36] Chaos Computer Club, 2022: "Chaos Computer Club hacks Video-Ident".

[i.37] European Commission, 2021: "Proposal for a Regulation of the European parliament and of the council laying down Harmonised rules on artifical intelligence (Artificial Intelligence act) and amending certain union legislative acts".

[i.38] Alexandre Sablayrolles et al., 2020: "Radioactive data: tracing through training".

[i.39] Zen et al., 2019: "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech".

[i.40] Kim et al., 2022: "Guided-TTS 2: A Diffusion Model for High-quality Adaptive Text-to-Speech with Untranscribed Data".

[i.41] Watanabe et al., 2018: "ESPnet: End-to-End Speech Processing Toolkit".

[i.42] Hayashi et al., 2020: "Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit".

[i.43] Chen et al., 2022: "Streaming Voice Conversion Via Intermediate Bottleneck Features And Non-streaming Teacher Guidance".

[i.44] Ronssin et al., 2021: "AC-VC: Non-parallel Low Latency Phonetic Posteriorgrams Based Voice Conversion".

[i.45] Tan et al., 2022: "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality".

[i.46] Liu et al., 2022: "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild".

[i.47] Müller et al., 2021, ASVspoof 2021: "Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn?".

[i.48] Müller et al., 2022, ASVspoof 2021: "Does Audio Deepfake Detection Generalize?".

[i.49] Gölge Eren, 2021: "Coqui TTS - A deep learning toolkit for Text-to-Speech, battle-tested in research and production".

[i.50] Min et al., 2021, Meta-StyleSpeech: "Multi-Speaker Adaptive Text-to-Speech Generation".

[i.51] Keith Ito, Linda Johnson, 2017: "The LJ Speech Dataset".

[i.52] Ganesh Jawahar, Muhammad Abdul-Mageed, Laks V. S. Lakshmanan, 2020: "Automatic Detection of Machine Generated Text: A Critical Survey".

[i.53] Rowan Zellers et al., 2019: "Defending Against Neural Fake News", Advances in Neural Information Processing Systems, Curran Associates, Inc.

[i.54] Original Deepfake Code, 2017.

[i.55] Matt Tora, Bryan Lyon, Kyle Vrooman, 2018: "Faceswap".

[i.56] Ivan Perov et al., 2020: "DeepFaceLab: A simple, flexible and extensible face swapping framework".

[i.57] Yuval Nirkin et al., 2019: "FSGAN: Subject Agnostic Face Swapping and Reenactment".

[i.58] Lingzhi Li et al., 2020: "FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping".

[i.59] Renwang Chen et al., 2021: "SimSwap: An Efficient Framework for High Fidelity Face Swapping".

[i.60] Jiankang Deng et al., 2018: "ArcFace: Additive Angular Margin Loss for Deep Face Recognition".

[i.61] Aliaksandr Siarohin et al., 2020: "First Order Motion Model for Image Animation".

[i.62] Justus Thies et al., 2020: "Face2Face: Real-time Face Capture and Reenactment of RGB Videos".

[i.63] Guy Gafni et al., 2021: "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Recontruction".

[i.64] Andreas Rössler et al., 2019: "FaceForensics++: Learning to Detect Manipulated Facial Images".

[i.65] TheVerge, 2021: "Tom Cruise deepfake creator says public shouldn't be worried about 'one-click fakes'".

[i.66] Matt Tora, 2019: "[Guide] Training in Faceswap".

[i.67] J. Naruniec et al., 2020: "High-Resolution Neural Face Swapping for Visual Effects".

[i.68] H. Khalid et al., 2021: "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset".

[i.69] W. Paier et al., 2021: "Example-Based Facial Animation of Virtual Reality Avatars Using Auto-Regressive Neural Networks".

[i.70] L. Ouyang et al., 2022: "Training language models to follow instructions with human feedback" (GPT35).

[i.71] P. Christiano et al., 2017: "Deep reinforcement learning from human preferences" (RLHFOriginal).

[i.72] OpenAI, 2022: "Introducing ChatGPT" (ChatGPT).

[i.73] A. Glaese et al., 2022: "Improving alignment of dialogue agents via targeted human judgements" (Sparrow).

[i.74] J. Menick et al., 2022: "Teaching language models to support answers with verified quotes" (GopherCite).

[i.75] Emily M. Bender et al., 2021: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?".

[i.76] J. Devlin et al., 2019: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

[i.77] G. Lopez, 08.12.2022: "A Smarter Robot", The New York Times.

[i.78] P. Mukherjee et al., 2021: "Real-Time Natural Language Processing with BERT Using NVIDIA TensorRT (Updated)".

[i.79] F. Nonato de Paula and M. Balasubramaniam, 2021: "Achieve 12x higher throughput and lowest latency for PyTorch Natural Language Processing applications out-of-the-box on AWS Inferentia".

[i.80] F. Matern et al., 2019: "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations", IEEE™ Winter Applications of Computer Vision Workshops.

[i.81] A, Azmoodeh and Ali Dehghantanha, 2022: "Deep Fake Detection, Deterrence and Response: Challenges and Opportunities".

[i.82] N. Yu et al., 2021: "Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)IEEE™ International Conference on Computer Vision (ICCV).

[i.83] B. Guo et al., 2023: "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection".

[i.84] Insikt Group, 2023: "I, Chatbot", Recorded Future.

[i.85] Cade Metz, 2023: "OpenAI to Offer New Version of ChatGPT for a $20 Monthly Fee", NYT.

[i.86] Joseph Cox, 2023: "How I Broke Into a Bank Account With an AI-Generated Voice", Vice.

[i.87] C. Wang et al., 2023: "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers".

[i.88] Coalition for Content Provenance and Authenticity, 2023: "Overview".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

deepfake: manipulation of existing or creation of fake multimedia identity representation

face reenactment: method for creating deepfakes in which the facial expressions of a person in an video are changed

face swap: method for creating deepfakes in which the face of a person in an video is exchanged

multimedia identity representation: data representing a person's identity or linked to it in different media formats such as video, audio and text

Text-To-Speech (TSS): method for creating deepfakes in which text (or a phoneme sequence) is converted into an audio signal

voice conversion: method for creating deepfakes in which the style of an audio sequence (e.g. speaker characteristic) is changed without altering its semantic content

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AI	Artificial Intelligence
AML	Anti-Money Laundering
API	Application Programming Interface
BEC	Business E-mail Compromise
CEO	Chief Executive Officer
DNN	Deep Neural Network
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HTML	Hyper Text Markup Language
ID	Identity
KYC	Know Your Customer
MOS	Mean Opinion Score
NLP	Natural Language Processing
RLHF	Reinforcement Learning from Human Feedback
TTS	Text-To-Speech
VC	Voice Conversion

4 Introduction

4.1 Problem Statement

The present document covers the AI-based manipulation of multimedia identity representations. Due to significant progress in applying AI to the problem of generating or modifying data represented in different media formats (in particular, audio, video and text), new threats have emerged that can lead to substantial risks in various settings ranging from personal defamation and opening bank accounts using false identities (by attacks on biometric authentication procedures) to campaigns for influencing public opinion. AI techniques can be used to manipulate authentic multimedia identity representations or to create fake ones. The possible output of such manipulations includes, among other things, video or audio files that show people doing or saying things they never did or said in reality. Since usually Deep Neural Networks (DNNs) are used for generating such outputs, they are commonly referred to as "deepfakes".

In principle, this phenomenon is not entirely new, since somewhat similar attacks have by now been possible for an extended period of time. Falsely associating people with text they have never uttered does not require complex technology and has been done for millennia. Similarly, photos, audio and video files can be used out of their original context and attributed to a completely different one. Although this technique is very unsophisticated, it can be remarkably successful, and is still routinely used, e.g. in today's social networks. The rapid advance of computer technology in recent decades also made the manipulation of photos, audio and video files increasingly easier. Editing programs allow cropping and rearranging audio and video files or changing their speed. Since photo-editing programs became widespread in the 2000s, the possibilities for manipulating photos have been practically unlimited.

EXAMPLE: In 2020, a video showing US Speaker of the House Nancy Pelosi circulated on social media. The video had been slowed down to give the impression of Mrs. Pelosi being drunk [i.1].

Nevertheless, AI techniques allow going one step further in many respects and can have adverse effects in a larger array of situations. AI techniques allow automating manipulations that previously required a substantial amount of manual work, creating fake multimedia data from scratch and manipulating audio and video files in a targeted way while preserving high acoustic and visual quality of the result, which was infeasible using previous technology. AI techniques can also be used to manipulate audio and video files in a broader sense, e.g. by applying changes to the visual or acoustic background. However, such manipulations do not target the identity representations of the persons involved. The present document focuses on the use of AI for manipulating multimedia identity representations and illustrates the consequential risks and measures to mitigate them.

5 Deepfake methods

5.1 Video

5.1.1 General

This clause discusses the methods available for the manipulation of image sequences from video data. The audio part of video data is discussed separately within clause 5.2, as well as the combination of manipulated image sequences with audio data in clause 5.4. Multiple methods based on deep neural networks exist for the editing of image sequences. These methods were developed for achieving various objectives. They include methods for "face swapping" and "face reenactment" / "puppeteering". Beyond face swapping and reenactment, further AI-assisted video editing methods are available or actively researched, but not yet as popular. Full-body puppeteering [i.13] methods aim to transfer the body movement of a person to another person. In addition to the aforementioned methods, which generally use identity attributes from another existing person to perform the manipulation of image sequences, fully synthetic data can also be created.

5.1.2 Face swapping

Face swapping is possibly the most famous method in social media and the general public, and also the one which coined the term "deepfake". The term became popular in 2017 when a user with the pseudonym "deepfakes" started to insert faces of celebrities into pornographic material using a neural network as an autoencoder model and posted the results on the web platform reddit [i.10]. The aim in face swapping is to change the identity of a person by changing either the core part of the face or the entire head. In this context, the neural network is trained to extract relevant information such as the face identity, expression and lighting conditions from an input image, and to generate a facial image of the target identity with the same expression and lighting conditions for seamless insertion into the frame.

The purpose of a face swap can be either entertainment, for example when inserting a popular celebrity's face into a movie scene that he/she originally did not participate in, or nefarious activities as in the case of non-consensual pornography (for details see clause 6.1.2). It can also be used for other purposes, as for a more natural de-identification (opposed to face blurring) within a documentary film. This allows keeping the respective persons' emotional expressions but protects them from prosecution [i.11].

5.1.3 Face reenactment

If one does not aim to manipulate the identity of a speaker but for example to alter a spoken message, face reenactment methods can be used for editing a given video.

EXAMPLE: In an early video from 2018 former president of the USA Barack Obama warns of an upcoming era of disinformation and insults acting president Donald Trump, just to reveal afterwards that the video was manipulated all along [i.12].

As the identity of the person in the video is preserved in this method, only subtle changes need to be made in the facial expression or in the region of the mouth. This manipulated content can then be inserted seamlessly, and can achieve higher quality in comparison to face swapping methods as differences in skin color or texture do not need to be considered. However, the general setting of the video is mostly determined by the original source material that is being manipulated, unless further manipulation steps are applied to the body of the manipulated person or the background.

5.1.4 Synthetic faces

Using techniques such as StyleGAN2 [i.2], it is possible to create 2D pictures of synthetic faces at a resolution of 1024x1024 pixel, which show faces of people that might not exist in reality. On the technical level, the goal of these systems is usually to map a simple random distribution, such as a multivariate Gaussian distribution, onto the distribution of natural faces. For creating a new face, a vectoris first sampled from the simple distribution, which is then converted by the system into a two-dimensional image. The mapping of the two distributions is generally modelled with a deep neural network. Usually Generative Adversarial Networks (GANs) or Variational Autoencoders are used for this task.

Modern methods based on this technology are also capable of creating three-dimensional representations of random pseudo-identities [i.3]. Furthermore, these systems can also be used to manipulate facial attributes of the created faces. The input vector or an intermediate representation of it is often changed in a controlled manner, which results in the change of the specified attribute in the output of the system. In some cases, however, methods for changing attributes still have the problem that other attributes are also changed during this process.

EXAMPLE: The age, facial expression, or hair color of a pseudo-identity can be controlled and manipulated using StyleFlow [i.4]. However, if an attribute is changed too much, it can have the side effect of changing the interpretation of gender of the person, for example.

In addition, those systems also provide the ability to generate facial images of real people, whose attributes can in turn be manipulated [i.5]. The ability to manipulate real faces by means of these methods even allows the morphing of several faces into one face, which contains biometric characteristics of all the original faces [i.6].

On the one hand, synthetic faces can be used by attackers to conceal their identity or to create fake profiles on social media in the scope of disinformation operations. On the other hand, synthetic faces can also be used for anonymization for legitimate purposes.

5.2 Audio

Methods for the creation of manipulated audio data have the goal of creating audio data that contain a given semantic content and have a specified style.

This class of manipulation methods can be divided into two main categories, Text-To-Speech (TTS) methods, which can be used to generate synthetic audio data and Voice Conversion (VC) methods, which can be used to manipulate existing audio data.

Text-To-Speech methods can be used to convert a certain semantic content, which is specified by a text or a phoneme sequence, into an audio signal. The generated audio signal should contain the specified semantic content and be perceived to be as natural as possible by a human listener [i.7].

Frequently, TTS methods also have the option of controlling the style of the generated audio signal. This can be used, for example, to control the speaker, the emotion, or the speech rate of the audio signal. Modern TTS methods are usually designed as multi-speaker systems, which makes it possible to define the speaker whose characteristics are to be included in the generated audio signal at inference time. In some cases, it is also possible to generate forgeries of speakers who were not present during the training phase of the TTS method by providing the TTS system with real audio material as a reference at the inference phase ("one-shot") [i.9]. However, if high-quality fakes which approximate the speaker characteristics of the target speaker as well as possible are to be generated, it is necessary that data on the target speaker is contained in the training set of the system.

Usually, a lot of audio data and the corresponding transcription are needed to train such models. Furthermore, in addition to multi-speaker methods, there are also multi-language methods, which make it possible to specify at system runtime which language the given text is. This makes it possible to achieve better results for languages for which only few training data are available. Most TTS methods consist of two components, a "text-to-spectrogram" module and a vocoder, which are usually both modeled with the help of deep neural networks.

The former is used to convert a text, or other representation of semantic content, into a lossy spectral representation, which is usually a mel spectrogram. The vocoder, on the other hand, is used to generate an audio signal from this representation.

Voice conversion techniques can be used to convert a source audio signal into another audio signal in such a way that the semantic content remains, but the style of the audio is changed according to the given specification. Such style changes could be a change of the speaker characteristic, a change of emotion, or a change of speech rate.

EXAMPLE: The most common application of voice conversion methods is to convert one audio file into a new file by changing the voice of the source speaker to a specified target speaker. The output audio file contains the same semantic content as the source audio but sounds like the target speaker's voice.

In addition to the two common components ("text-to-spectrogram" and vocoder) used in TTS systems, VC methods usually have a component that decomposes the source audio signal into different representations, such as the semantic content, timbre, or prosody [i.8].

5.3 Text

In the past years, the area of Natural Language Processing (NLP) has evolved steadily. NLP includes several tasks like question-answering, machine translation, summarization and also text generation. Due to the success of several so called language models (roughly speaking, models that are trained to predict the likelihood of a word or sentence, given a context), NLP is receiving increasing attention from scientists as well as the public [i.27]. There is no clear definition of deepfakes in the text domain; however, in the present document the term "deepfake" is used when a text is machine-generated with the intention to appear human and to spoof an entity (e.g. a specific person, company or organization). Moreover, the term is mostly used in the context of targeted or untargeted deceptive attacks. Other possibilities of malicious use of language models, e.g. polymorphic malware generation [i.35], also exist. They are out of scope for the present document. The following text is focused only on the threats posed by automatically generated human-like text with the intention to spoof an entity.

Starting a few years back, concerns were growing that language models could be misused in order to either harm people with fraudulent texts (e.g. phishing, spam or CEO -fraud) or to fool people or society at large by generating misleading or fake content (e.g. fake news). Besides that, NLP models provide further use cases to deceive human individuals. Recently published models provide the ability to generate literature or scientific research papers raising the question of responsibility for the content or legal issues in terms of authorship and copyright [i.31].

Due to their ability of writing highly convincing human-like texts, several tech companies prevented, limited or delayed the access to or the release of their models to impede misuse [i.33], [i.26] and [i.23]. In order to be able to understand how these models work and why they are well performing in various text-based tasks, the next paragraph gives a brief theoretical overview.

Most state-of-the-art large language models are based on the transformer architecture that was presented in 2017 by Vaswani et al. [i.22]. Transformer models are using word embeddings combined with a distinct positional encoding as input. Word embeddings are a vector-based representation of words, whereas the positional encoding contains information about the position of each word within the input. The original transformer architecture presented in 2017 consists of an encoding and a decoding block. The main intention behind this architecture is to reduce the input within the encoder to a lower- dimensional space (e.g. reducing a word to its meaning) and reconstruct it via the decoder (e.g. translation into a different language). Transformer models differ from former language models in the use of so-called self-attention as their core architecture. This self-attention mechanism represents the relationship between each word of the input-text and every other word within the text [i.22].

There are various types of transformer models, two of which are often discussed in the context of generation and detection of fake content. Bidirectional language models, on the one hand, are transformers consisting of the encoder part only. Among other things, this transformer architecture shows good results in question-answering or in detecting certain automatically generated texts [i.23], which will be further discussed in clause 8. On the other hand, unidirectional transformer architectures are solely based on the decoder module of the original transformer presented in [i.22]. They process the text from left to right and have to predict the next word. Therefore, they are extraordinarily good in generating texts [i.30].

Training transformer models contains an unsupervised or self-supervised pre-training step with unlabeled data. After that, the model can either be fine-tuned for a specific task (which can make it less universal but well suited for the trained use-case), or directly used via zero-shot transfer, one-shot or few-shot learning. To use the model directly, it is sufficient to provide it with a description of a task, written in natural language, followed by either no (zero-shot), one or few examples. This makes these models extremely easy to use. The authors of [i.32] state that when the size of the model is large, increasing the number of shots will increase the precision to a level similar to that of fine-tuned models. More recent language models like OpenAI's ChatGPT [i.72] or DeepMind's Sparrow [i.73] and GopherCite [i.74] combine fine-tuned transformer models with reinforcement learning for parameter optimization. The algorithm is called Reinforcement Learning from Human Feedback (RLHF) and is based on an idea that was already presented in 2017 [i.71]. This optimization method leads transformer models to generate more aligned and human-like texts.

5.4 Combinations

Deepfakes that make use of both audio and image sequences to create a manipulated video are mostly used in the context of either video dubbing to achieve lip synchronicity or for the audio-driven generation of talking heads.

A popular method to lipsync a video with the original image sequence to a given audio file with human speech is to use a neural network to detect and rerender the mouth region in the images [i.15]. Alternative to the audio input, the manipulation of the mouth region can also be done by only editing a few words via text input [i.16]. Such methods can be used in combination with face swap and voice conversion or Text To Speech methods for a combined swap of facial and speech identity in a given video, as demonstrated in [i.68].

For the generation of a talking head, one method in a GAN based setup is to use an image of a person as identity reference, an image sequence as driver for the head pose, and an audio sequence as driver for the shape of the mouth region [i.17]. By this approach the pose of the head and the mouth movement are separated. However, as the target identity is only generated from a single reference image, biometric correctness of the face from varying view angles is not to be expected. Another approach to a talking head is to generate animatable three-dimensional face models of a target identity, which is then animated and rendered according to a text or audio sequence [i.69]. As shown in [i.69], the model can be created by photogrammetric means using multiple cameras simultaneously on a person from different view angles and the model can then be used to animate the pronunciation of words that are not part of the training data.

The combination of audio, video and text deepfakes of a person allows a novel kind of " duplication attack" in which an arbitrary number of false "copies" of a person can stage an attack at a given time.

EXAMPLE: A large number of deepfake copies of a politician could make calls to other politicians, thus increasing the chance that the fake is not noticed in time by all interlocutors.

6 Attack scenarios

6.1 Attacks on media and societal perception

6.1.1 Influencing public opinion

Methods for manipulating multimedia identity representations can be used to exert influence on public opinion. The modus operandi consists in publishing manipulated media which falsely create the impression that persons holding influential positions have written, said or done certain things. In principle, this applies to all purposes where stakes are high and the benefit justifies the effort from an attacker's perspective.

From an economic perspective, skilfully deployed deepfakes can be used to manipulate prices on the capital markets, either for the benefit or to the detriment of the targeted stocks or goods. As these prices may change very quickly, such an approach might be valuable since an attacker could benefit from price changes even if the deepfakes were exposed and prices went back to normal after a short time. In a related attack, a company could try to disparage competitors. For this to have a lasting effect would however require to sustain the effort for some time. Fake product reviews constitute a specific attack vector from an economic perspective. Their aim is to promote a specific product in a positive way or to disparage another one, which can influence sales numbers.

In political terms, a valuable objective is to sway public opinion in the run-up to elections or plebiscites. History is rife with incidents where public opinion before elections was strongly influenced by single events, and an attacker may try to benefit from this volatility. An attacker may likewise try to slowly alter people's opinion by repeatedly publishing deepfakes discrediting one of the competing parties. An attacker can use fake video or audio content to this effect. Recent language models such as GPT-3 or ChatGPT can also be used for accelerating and partly automating the creation of convincing fake news content [i.28], [i.84]. This content imitates the writing style of real news media content and can aim to spread disinformation [i.34]. Language models can also be used for astroturfing operations, which usually operate on social media and create the impression that a large crowd believes in or supports a specific topic. Thus, astroturfing imitates a so called grassroot movement and can be an efficient means to strengthen the reach of a disinformation campaign or conceal its origin [i.29].

Another use of deepfakes is as a propaganda tool, especially in times of war. Such incidents are increasingly being reported.

EXAMPLE: In March 2022, a deepfake video of Ukrainian president Volodymyr Zelenskyy was posted on social media platforms. In the video, Mr. Zelenskyy purportedly announced capitulation in the face of the Russian invasion of Ukraine. The video was quickly identified as a low-quality deepfake but likewise described as a harbinger of the future use of sophisticated deepfakes for disinformation purposes [i.19].

Deepfakes may be used to complement and enhance more traditional, manual methods for spreading disinformation. When assessing the threat level posed by deepfakes in this respect, one needs to take into account the specific side conditions, however. Indeed, [i.18] argues that propagandists may in many cases consider traditional manual editing of media files and using them out of context to have a higher return on investment, as exemplified by the disinformation campaign prior to the 2016 US presidential elections, where the attackers focused on scale rather than quality.

6.1.2 Personal defamation

Deepfakes may also be used for attacks on the personal level. In many cases, attacker and victim know each other and the attack is motivated psychologically or emotionally. These attacks usually consist in spreading fake video, audio or text on social media to the victim's peer group to ruin the victim's reputation or to humiliate them.

EXAMPLE: Particularly widespread is the creation of faked sexually explicit videos (often called revenge porn) where the victim's face is inserted into the original footage. This attack is mostly targeted against women [i.21].

The attack may also be targeted against public figures or celebrities. In this case, it may also have political or financial components.

6.2 Attacks on authenticity

6.2.1 Attacking biometric authentication methods

Another attack type explicitly targets procedures for remote biometric identification and authentication. Such procedures are in widespread use in many countries to give users access to digital services provided by public entities or private entities from regulated industry sectors.

Procedures for remote identification of a person via video are used in many European countries as a means for customers to open bank accounts. They cut costs for the financial industry compared to identification in situ while at the same time facilitating compliance with current Know Your Customer (KYC) and Anti-Money Laundering (AML) regulation. Speaker recognition systems are also used in the financial sector to authenticate customers requesting transactions.

The security level of these procedures and their resulting susceptibility to attacks using manipulated identity representations varies significantly. Whereas initially such identification procedures were based on photos provided by the users, most current methods rely on video data instead. Indeed, photo editing tools allow easily defeating the former procedures. The latter ones may be attacked using deepfakes, where the necessary effort depends on the specific side conditions and countermeasures in place (see clause 8.2.3). Attacks may involve biometric data from third persons obtained without their knowledge, or rely on purely synthetic data.

EXAMPLE: In 2022, the German hacker association Chaos Computer Club mounted successful attacks on video identification procedures by applying deepfake methods to ID documents [i.36].

6.2.2 Social Engineering

In many situations, biometric authentication is not carried out explicitly but is implicitly and unconsciously relied upon within human interactions. Attackers can target these implicit authentication procedures for enhancing the probability of success of social engineering attacks. Manipulated multimedia identity representations allow for much more convincing scenarios that can involve the writing style, voice or videos of persons supposedly communicating with the attack victims. In general, these attacks can be seen as an upgrade to previous, less sophisticated attacks involving facial masks, doppelgangers or voice mimicking.

Generic phishing attacks and spear-phishing attacks, which target specific recipients, in most cases aim to have their victims click on malicious links, which attackers can then either use to obtain login credentials or to distribute malware.

These attacks are usually text-based and can be made more convincing and customisable by relying on recent language models [i.25].

Another type of social engineering that is suitable for deepfake-based enhancement is the CEO fraud attack (also called Business E-mail Compromise (BEC)) [i.24], in which usually an attacker impersonates a senior person from an organization contacting one of the organization's employees and requesting them to transfer a substantial amount of money to an account controlled by the attacker.

EXAMPLE: In 2020, a bank manager in Hong Kong was fooled by attackers that faked the voice of a company director. The criminals' booty amounted to $35 million [i.20].

6.3 Digression: Benign use of deepfakes

Apart from manifold attacks, deepfakes can also be put to benign uses. In the artistic area, they can be used to create movie content that could otherwise not be created because the actors involved have deceased or their outward appearance has considerably changed because of ageing or other reasons. They can also allow improving on traditional anonymisation techniques used in documentaries, which usually consist in blurring the proteted person's biometric characteristics. Deepfakes can help to better convey the person's emotions by keeping their facial movements or prosody while still protecting their anonymity.

In the area of biometrics, synthetic data can help to comply with data protection regulation such as the European GDPR, which includes strict requirements when dealing with people's biometric data. This makes it quite challenging to assemble large-scale data bases that are required for training biometric systems to make them achieve good performance. Synthetic data may help to mitigate this problem.

7 State of the art

7.1 Data

7.1.1 Data required for Video Manipulation

The data required for effective face manipulation highly depends on the respective tools used and the objective that an attacker wants to achieve. Current face swapping tools that need to learn to encode and decode the specific identity of a person generally need a video that shows a person from various perspectives and need an order of at least 500 - 1 000 frames to produce high-quality face swaps [i.66]. The needed amount also depends on the scene that the face swap is merged into. If varying light conditions and extreme facial positions need to be rendered, more specific training material for such conditions is necessary. To cover most possible cases that occur during a faceswap, the data should be diverse in viewing angle, lighting conditions and facial expressions [i.66]. The image resolution of the training data should also not be much lower for the target face to avoid a notable jump in image quality in the face area. Some autoencoder models, such as [i.56], allow using models that are pretrained on large facial image data sets to decrease training time by transfer learning. A caveat of these specific models is, however, identity bleeding, where generated faces contain features from the input face and the target face, as they use a shared decoder [i.66]. Other approaches, such as [i.67], show that it is also possible to train only the encoder on multiple identities, and have a single decoder for each target identity. In their work, the authors observe an improvement in expressability of the generated facial images in comparison to using only two identities, but mention as trade-off an increase in training time linear in the number of identities trained simultaneously.

On the other hand, methods exist that are trained in a subject-agnostic architecture on a large face database and that need only a single image of a target identity to extract an identity representation (see clause 7.2.1 for more details on these tools). Good results can also be achieved in this case, but especially when the head turns and moves heavily, the lack of information resulting from using only a single image can result in visible inconsistencies.

7.1.2 Data required for Audio Manipulation

Creating manipulated audio content of a target speaker requires audio material of this person speaking. In order to reduce the amount of data required, multi-speaker systems are usually used, which are able to generate an audio signal for the target speaker on the basis of a speaker embedding. For this purpose, a multi-speaker system is usually pre-trained on large data sets (e.g. [i.39]) and later finetuned for the target speaker. This significantly reduces the amount of data needed from the target speaker. In current research results, a few minutes to a few seconds of audio material from the target speaker are sufficient [i.9], [i.50], [i.40] and [i.87]. Another important class are so-called one-shot methods, which only need a few seconds of audio material of the target speaker in the inference phase as a reference, i.e. they are not trained with the data of the target speaker at all. However, these methods usually work worse than those in which the model is trained with the target speaker's data [i.40].

7.1.3 Data required for Text Manipulation

Until now, the quality of texts generated by language models still depends on the amount and quality of available data. Current models are based on gigabytes of text from a wide variety of sources like Reddit, Wikipedia, book corpora etc. up to most of the internet [i.75]. This huge amount of text is necessary to calculate the word probabilities for the most likely next word of a text as accurately as possible. As a result, language models make far fewer errors in those languages for which they were trained than they did a few years ago [i.75].

In general, depending on the architecture of the model, there are different kinds of text training data for transformers. When training a unidirectional transformer, the data for training is usually preprocessed (cleaned of e.g. HTML tags and converted into word embeddings) and then passed through the transformer word by word. The transformer generates the next word depending on the input it received. Bidirectional transformers are trained with (also preprocessed) but masked texts. The masks are gaps within the text which make up about 15 % of the words. Among other things, bidirectional models are trained by filling in words in the gaps in order to reconstruct the original text instead of predicting the next word of an input sequence [i.76].

New architectures like transformers based on the RLHF approach further optimize the alignment of a language model's outcome. In other words, the text a language model generates matches the expectations of users more precisely and sounds more human-like. This is achieved by additional data sets that are carefully created by the developers. These data sets contain prompts of users or so-called labellers together with desired model outputs. This data is very cost-intensive to create as it needs a lot of human resources [i.70]. However, the outcome is much better than the results obtained by traditional methods, as many news papers, scientists or experiments by users on social media show [i.77].

7.2 Tools

7.2.1 Tools for Video Manipulation

For the purpose of face swapping and reenactment, a multitude of tools have been developed within the last five years and new ones are being developed still. The tools mentioned below are therefore only a selection of those tools that are currently (2023) most popular or often being used in research, for example in large deepfake databases.

From the source code that was used to generate the material published on the r/deepfake subreddit in 2017 [i.54], two major open source projects were created that further improved upon the initial work: FaceSwap.dev [i.55] and DeepFaceLab [i.56]. Whereas the first project states that it is developing tools for ethical usage of deepfake videos, the latter also provides a platform for sharing pornographic material and openly shares models for said usage. The project FaceSwap.dev is only focused on the offline creation of deepfakes with no code for real-time application. DeepFaceLab, on the other hand, mantains a project called DeepFaceLive to apply trained models in real time on a webcam feed for example. Extensive guides on how to use both projects are available in the respective forums and large active communities have formed around both.

The models in the aforementioned faceswap frameworks need to be trained with the specific identities that are swapped. Other approaches for face swapping have also been published that are subject-agnostic and do not need to be specially trained, where among the most prominent are FSGAN [i.57], FaceShifter [i.58] and SimSwap [i.59]. FSGAN follows a different architecture from the autoencoder and generates the faces by multiple GAN-trained networks, whereas the latter two models are broadly speaking autoencoder models. They additionally use a face recognition model, ArcFace [i.60], to extract an identity feature vector and use it in the decoding step for the generation of the face swap. However, high-resolution results of face swaps that are shared in media as state-of-the-art face swaps are generally still performed with networks that are specifically trained on the respective identities.

For face reenactment, models exist that allow transferring the facial expressions from a driving video onto a single image (e.g. First Order Motion Model [i.61]), a video (Face2Face [i.62]), or onto a facial avatar (e.g. [i.63]) of the target identity. In many cases, except for the reenactment of single images, these tools are currently closed source and only available for research through prepared videos in deepfake detection databases such as FaceForensics++ [i.64].

Another broad class of tools for the creation of video manipulations are mobile applications such as Reface, FaceApp or Avatarify. These applications are either cloud applications that process an input image or video for the generation or manipulation of facial images, or, in the case of Avatarify, they are an implementation of a popular open-source project that can run on a mobile phone and has an improved user interface (First Order Motion Model in this case).

In general, modern video-editing tools such as Adobe After Effects implement neural networks in their toolbox to allow for quick image segmentation [i.14] in a video.

7.2.2 Tools for Audio Manipulation

There are several public tools for creating or manipulating synthetic voices. On the one hand, there are free open source frameworks in which different TTS and VC methods are implemented [i.41], [i.42] and [i.49]. However, these usually require a familiarization phase for the user and powerful hardware for the training and execution of the models.

On the other hand, there are cloud-based solutions for VC and TTS methods, some of which can also be extended with their own speakers, which usually require less user expertise and no powerful hardware.

In addition, the technology is becoming increasingly accessible to the public.

EXAMPLE: With the cloud service ElevenLabs, even laypersons are able to clone the voice of any person, requiring only a few seconds to minutes of audio material of that person. Using synthetic audio created with this service, a journalist was able to bypass the automated speaker recognition of a British bank [i.86].

7.2.3 Tools for Text Manipulation

After developing the first well-performing transformer models like GPT-2, researchers and companies were afraid of their potential dangers and made them available only to selected users, if at all. However, after a few months some models were made availableto the general public [i.23].

As of now, several language models are available for paying customers or even for free. The cost for using a paid model is about a few cents per thousands of tokens. Thus, even language models that do cost money are comparatively affordable [i.79]. Using the free version often implies that the data or texts entered into the model API can be further used by the company for improving its models. As a result, the access to language models is comparatively easy and often requires only the creation of a user account.

7.3 Latency

7.3.1 Latency in Video Manipulation

As most face swap and reenactment methods work on a video frame by frame, the latency is determined by the computation time needed for manipulating a single frame. Multiple steps often have to be performed on an input frame to generate a manipulated output frame, such as: face detection, landmark detection, swap face generation, skin color adjustment and blending of the manipulated material into the original frame. For decent quality manipulations with face resolutions in the order of 250 pixels, latencies in the order of 10 - 100 milliseconds are often observed for tools such as [i.56]. As this is, however, in the order of internet latency, facial swaps are generally hard to spot by their latency.

7.3.2 Latency in Audio Manipulation

Most voice conversion methods were not designed in an autoregressive manner for the purpose of real-time conversion, where only the current chunk of recorded audio data and the chunks from the past can be used for the continuous conversion process. Instead, they also use chunks of data following the time frame to be converted. However, there are also developments in research which are specially designed for real-time conversion and therefore manage with a low delay of approximatively 0,3 seconds [i.44] and [i.43]. Additionaly, it needs to be noted that those works mainly focus on converting the timbre of the voice in the audio and less on other factors like the prosody.

7.3.3 Latency in Text Manipulation

Language models that are made public by big companies via online access like ChatGPT are capable of generating texts in a few seconds after the command is typed [i.78] and [i.79]. After that, they continue generating text as long as needed or a maximum length of tokens is reached. Generating long texts automatically takes a fraction of the time a human would need to write the same text from scratch. However, due to the popularity of some language models, access may not always be possible due to limited server capacity. To overcome this problem, access via a paid version allows users to access the service at any time [i.85].

7.4 Distinguishability

7.4.1 Distinguishability of Video Manipulation

Video manipulations that contain swapped faces are becoming harder to distinguish by eye from authentic data. Usual telltale signs can be searched at the transition between unmanipulated areas of the image and the inserted face. Depending on the quality of the material and the post-processing that has been performed these can, however, be made quite subtle as prominent examples in social media already demonstrate [i.65]. For face reenactment, the manipulation is often even harder to spot, as the identity of the face is preserved and only the facial expressions are changed, which allows keeping transitions in skin texture and color to a minimum.

7.4.2 Distinguishability of Audio Manipulation

In the context of synthetic voices, the quality of the generated audio data is generally evaluated in terms of the Mean Opinion Score (MOS), which is the arithmetic mean of individual human ratings of the perceived quality of the audio signal (on a range of 1 - 5). The gap between synthetic audio data and real data highly depends on the type of data.

EXAMPLE: In the case of the LJSpeech dataset [i.51], which comprises approximately 24 hours of a single speaker reading passages from several books, current models have already achieved the same MOS value compared to the the real data [i.45]. For more difficult tasks, such as the generation of longer passages of a conversation, there is still a gap between real and synthetic data.

7.4.3 Distinguishability of Text Manipulation

The ability of humans to detect whether a text was generated by a language model or not depends on their experience with the language model, the topic and language of the text and the method of the detection task [i.83]. In general, human distinguishability is expected to deteriorate as the performance of language models improves.

EXAMPLE: The developers of GPT-3 found that the accuracy of humans in detecting texts generated by their biggest model (175 billion parameters) is only 52 % [i.32]. This finding supports the results presented in [i.25] where scientists analysed the effectiveness of AI used to automatically generate spear-phishing messages compared to that of human writers. They figured out that targets are more likely to click on phishing links within AI generated e-mails than on links within human-written e-mails.

8 Countermeasures

8.1 General countermeasures

Countermeasures against the manipulation of multimedia identity representations can be grouped into general countermeasures, which can be applied regardless of the specific attack scenario, and specific ones that take the side conditions and the context of attacks into account. Deepfakes pose a complex problem, for which there is no panacea but which can best be combated by a combination of measures on various levels.

A generic measure to mitigate the threat of deepfakes is to promote education and to raise awareness about the existence of the phenomenon and about what is currently possible, which goes well beyond what many people think, and what are the current limitations of deepfakes that allow discerning them. These educational measures can be both generic or more tailored to specific attack scenarios, e.g. by alerting users about potentially faked content on social media, or company employees about the options to enhance social engineering attacks. Users that are aware of these threats in the first place are much more likely to scrutinize media content and its origin, especially if the purported source, the context or subliminal clues suggest that there may be something wrong with it. Users can also be trained to pay special attention to aspects that currently can give away deepfakes, e.g. artifacts in the area of the teeth [i.80] in videos. However, such very specific pieces of advice may become quickly outdated when the state of the art advances.

A further generic countermeasure is given by regulatory interventions that require the use of manipulated identity representations to be clearly marked, as is foreseen by the proposed EU AI Act [i.37]. Admittedly, it is questionable whether this requirement will in itself mitigate the proliferation of deepfakes unless accompanied by good detection and enforcement mechanisms.

On the technical level, the problem can be countered by a range of detection methods. General media-forensic techniques can be used to analyse the content and look for artefacts introduced by imperfections in the deepfake generation methods. In video material, these artefacts may for instance be blurred areas, abrupt colour changes or pixel errors especially at transition points between the original footage and the synthetically generated parts. For texts, a common artefact is the repetitive usage of a relatively small set of words that are chosen by the language model with high probability [i.52].

Another approach is to use AI systems trained to detect manipulated content. Depending on the effort used for training these AI systems, they may work quite well, but it is likewise clear that attackers can enhance their attacks to elude existing detection methods, which gives rise to an arms race [i.81]. In addition, many methods seem to work only in a very limited laboratory scenario and generalize very poorly to realistic conditions, such as variation in recording device [i.46], [i.47] and [i.48], or only recognize content generated using a particular model [i.53].

A special variety of detection methods relies on distinctive patterns introduced into the training data used for creating the deepfakes. As such, it can be regarded as a poisoning attack that later on allows systematically recognizing fakes created using the poisoned data sets. Alternatively, companies that provide deepfake tools for legitimate purposes may also integrate such patterns directly into all media created by their tools. Different approaches including watermarks or so-called radioactive data exist [i.38], [i.82]. However, for these detection methods to have a chance of success, it is necessary that attackers use the respective data sets or tools and are unaware of this countermeasure.

8.2 Attack-specific countermeasures

8.2.1 Influencing public opinion

When manipulated multimedia identity representations are used with the aim to influence public opinion, which most often happens in social media, a specific countermeasure (besides the general ones mentioned in clause 8.1) is to ensure the authenticity and attributability of content. This can help users distinguish between content from confirmed trustworthy sources and content which originates from untrustworthy sources or for which no information whatsoever about the origin is available. A straightforward way to implement this is to use cryptographic protection mechanisms, in particular digital signatures. Media content originating from eminent public figures or official organizations could then be signed using these parties' private keys, and either the users themselves or the internet platforms distributing them could cryptographically check they have not been altered after the signature has been applied. If the respective keys can be securely bound to the respective parties, the authenticity of the information can likewise be checked. A notable project pursuing this approach is [i.88].

8.2.2 Social Engineering

The use of deepfakes for enhancing social engineering attacks can in principle be mitigated using the general measures from clause 8.1. When raising awareness among employees, the information can be specifically tailored to the predominant attacks. Employees can be trained to understand that both video and audio material can be quite convincingly faked and may be untrustworthy.

A more robust way to address the risk may be for companies and organizations to build robust processes, where important decisions and high-value transactions are not taken on the basis of implicit biometric authentication, but insteadare always confirmed using a standard procedure involving multi-factor authentication.

8.2.3 Attacks on authentication methods

Explicit attacks on (biometric) authentication methods can be addressed by increasing the difficulty of successfully creating fake content. The general strategy consists in introducing a high-level challenge-response protocol with the aim of producing easily discernible artefacts in fake content. In remote identification via video, the challenge-response protocol can require the person to be identified to perform specific movements, to move other objects through the image frame in a specific way or to produce specific occlusions and reflections. Speaker recognition methods can require the speaker to utter words that are hard to pronounce and with which audio generation methods struggle. Another strategy in this setting is to measure the delay in responses, since a large delay can be a hint that significant computational processing is happening. While this strategy can make successful attacks much harder, it is very likely that given sufficient effort and resources attackers will be able to circumvent it and produce convincing fakes also under the actions required by the challenge-response protocol.

Annex A:
Change history

Date	Version	Information about changes
04/2022	0.0.1	Fill skeleton with GR structure
05/2022	0.0.2	Add problem statement and methods for audio and synthetic faces
06/2022	0.0.3	Add methods for video and combination of modalities; add text on attack scenarios
07/2022	0.0.4	Added information on modality text
06/2023	1.1.1	First published version

History

Document history		
V1.1.1	June 2023	Publication