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 Reconstruction".

[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.